2028: Basic Statistical Methods
Homework 6: Solutions

1. **Simple linear regression**

11-3 We read the data and extract the response and predictor variables using the following R commands:

```
data = read.csv("data113.csv",header=TRUE)
y = data$Rating
x = data$Yds
```

To estimate the model parameter we need to fit a linear regression model. The R command is

```
model = lm(y~x)
summary(model)
```

The regression output is:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.195      9.059   1.567    0.128
x             10.092      1.288   7.836 9.59e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.219 on 30 degrees of freedom
Multiple R-squared: 0.6718,     Adjusted R-squared: 0.6609
F-statistic: 61.41 on 1 and 30 DF,  p-value: 9.589e-09
```

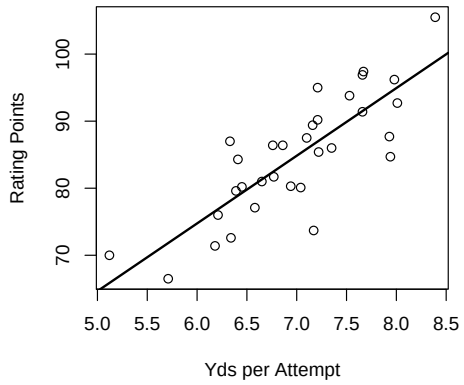
(a) From the R output, the model parameter estimates are as follows:

$$\widehat{\beta}_1 = 10.092, \widehat{\beta}_0 = 14.195, \widehat{\sigma}^2 = MSE = 27.24$$

To graph the regression line, we can use the following R commands

```
plot(x,y,xlab="Yds per Attempt",ylab="Rating Points")
abline(14.195,10.092,lwd=2)
```

The plot is in the figure below. From this plot, we find that there is a linear relationship between rating and the average number of yards gained per attempt.



(b) $\widehat{y} = 14.2 + 10.1(7.5) = 89.95$

(c) $-\widehat{\beta}_1 = -10.1$

(d) $\frac{1}{10.1} \times 10 = 0.99$

(e) $\widehat{y} = 14.2 + 10.1(7.21) = 87.02$
$e = y - \widehat{y} \Rightarrow e_1 = 90.2 - 87.02 = 3.18, e_2 = 95 - 87.02 = 7.98$

11-25 (a) We need to test: $H_0 : \ \beta_1 = 0$ vs. $H_A : \ \beta_1 \neq 0$. From the R output obtain in 11-3, we find that the t-value is $t = 7.836$ and the corresponding p-value is $9.59e-09$ (≈ 0). From this, we conclude that we reject the null hypothesis at $\alpha = 0.01$. Therefore, there is a significant linear relationship between rating and the average number of yards gained per attempt.

(b) The standard errors are: $se(\beta_0) = 9.059$ and $se(\beta_1) = 1.288$.

(c) The test is $H_0 : \ \beta_1 = 10$ vs. $H_A : \ \beta_1 \neq 10$. The test statistic is

$$t = \frac{\widehat{\beta}_1 - 10}{se(\beta_1)} = \frac{10.092 - 10}{1.288} = 0.07142857$$

The rejection rule is $|t| > t_{\alpha/2,30} = 2.749996$. Since the computed test statistic is smaller than $t_{\alpha/2,30} = 2.749996$ we conclude that we do not reject the null hypothesis at $\alpha = 0.01$.
Yes, the statement about the interpretation of the test is correct.

11-41 We estimate the 95% confidence intervals ($\alpha = 0.05$) as follows

(a)

$$\begin{aligned}\left(\widehat{\beta}_1 - t_{0.025,30} se(\widehat{\beta}_1), \widehat{\beta}_1 + t_{0.025,30} se(\widehat{\beta}_1)\right) &= (10.092 - 2.042 * 1.288, 10.092 + 2.042 * 1.288) \\ &= (7.461, 12.722)\end{aligned}$$

We conclude that the slope parameter β_1 is between 7.461 and 12.722 with a confidence level of 95%.

(b)

$$\left(\widehat{\beta}_0 t_{0.025,30} se(\widehat{\beta}_0), \widehat{\beta}_0 + t_{0.025,30} se(\widehat{\beta}_0)\right) = (14.195 - 2.042 * 9.059, 14.195 + 2.042 * 9.059)$$
$$= (-4.303, 32.693)$$

We note that the confidence interval contains the value zero suggesting that zero is a plausible value for β_0.

(c) $y^* \pm t_{0.025,30}\sqrt{MSE\left(\frac{1}{32} + \frac{(8-\bar{x})^2}{S_{xx}}\right)}$
$y^* = \widehat{\beta}_0 + 8\widehat{\beta}_1 = 14.195 + 8 * 10.092 = 94.931$
$MSE = 5.219^2 = 27.237$
$(8 - \bar{x})^2 = (8 - 6.997)^2 = 1.0043$
$S_{xx} = \frac{MSE}{se(\beta_1)^2} = \frac{27.237}{1.6589} = 16.418$
which give

$$94.931 \pm t_{0.025,30}\sqrt{27.237\left(\frac{1}{32} + \frac{1.0043}{16.418}\right)} = (91.69, 98.17)$$

(d) The prediction interval of the rating when the average yards per attempt is $x = 8.0$ is for $\beta_0 + 8\beta_1$. We note that $x = 8$ is not among the x values for which we observe the rating points and therefore, we will compute the prediction confidence interval given by
$y^* \pm t_{0.025,30}\sqrt{MSE\left(1 + \frac{1}{32} + \frac{(8-\bar{x})^2}{S_{xx}}\right)}$
$y^* = \widehat{\beta}_0 + 8\widehat{\beta}_1 = 14.195 + 8 * 10.092 = 94.931$
$MSE = 5.219^2 = 27.237$
$(8 - \bar{x})^2 = (8 - 6.997)^2 = 1.0043$
$S_{xx} = \frac{MSE}{se(\beta_1)^2} = \frac{27.237}{1.6589} = 16.418$
which give

$$94.931 \pm t_{0.025,30}\sqrt{27.237\left(1 + \frac{1}{32} + \frac{1.0043}{16.418}\right)} = (83.79, 106.07)$$

Therefore, the rating points range between 83.7 and 106 when the the average yards per attempt is $x = 8.0$

2. **Smoking and Cancer**

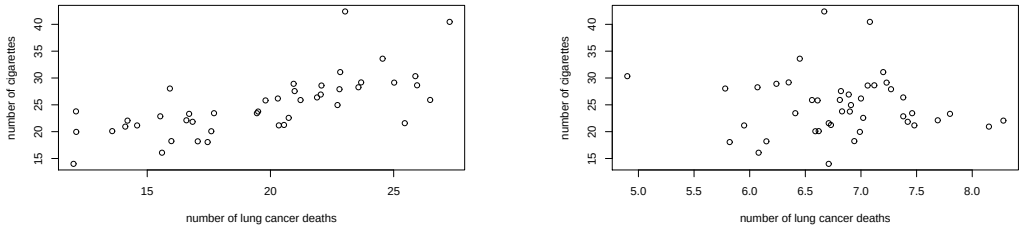
Question 1: Exploratory Data Analysis.

• From the two plots in Figure 1 we infer that there may be a significant linear upward association trend between the number of cigarettes smoked vs. the number of deaths from lung cancer but very weak association between the number of

cigarettes smoked vs. the number of deaths from leukemia (this is because the points are scattered randomly without a specific pattern).

• The correlation value between the number of cigarettes smoked vs. the number of deaths from lung cancer is 0.697 whereas the correlation value between the number of cigarettes smoked vs. the number of deaths from leukemia is -0.068. Similarly, we conclude that there is very low linear association between the number of cigarettes smoked vs. the number of deaths from leukemia.

• Yes it is reasonable to assume a simple linear regression model for the relationship between `CIG` and `LUNG` but not between `CIG` and `LEUK`.



(a) Lung Cancer (b) Leukemia

Figure 1: Scatter plots for assessing the linearity assumption.

Question 2: Fitting the Simple Linear Regression Model.

(i) The model parameters and their estimates are:

• The intercept parameter β_0 which is a scale parameter. For these data, its estimate is $\widehat{\beta}_0 = 6.4717$.

• The slope parameter β_1 which is indicates how strong the linear relationship between the two variables is. For these data, its estimate is $\widehat{\beta}_1 = 0.5291$.

• The error variance σ^2. For these data, its estimate is $\widehat{\sigma}^2 = 3.066^2 = 9.400356$.

(ii) The equation for the least squares line is No. of deaths $= 6.4717 + 0.5291 \times$No. of Cigarettes.

(iii) The slope parameter is positive and approximately equal to $0.5291(\pm 0.0839)$. Consequently, the higher the number of cigarettes smoked, the higher the number of deaths. In fact, the number of deaths increase with approximately 0.52 units at each additional 100 cigarettes smoked or one death per 100K for each additional 200 cigarettes smoked. This is a significant increase (note also that the p-value of the significance of the slope parameter is approximately 0).

(iv) A 95% confidence interval for the slope parameter is

$$(\widehat{\beta}_1 \pm t_{42,.025} se(\widehat{\beta}_1)) = (0.5291 \pm 2.018 * 0.0839) = (0.3598, 0.6984)$$

Question 3: Checking the Assumptions of the Model.

• Linear relationship between the two variables. This assumption holds as it is provided by the linear relationship displayed in the scatterplot in Figure 1.

• Normality of the data. We check this assumption by looking at the QQ normality plot and possibly the histogram of the residuals provided in Figure 2. Both plots indicate that the residuals are approximately normal (slight skewness to the left).

• Constant variance. We check this assumption by looking at the residual plot (upper right plot). The variance is constant as there is not a consistent pattern of the variability in the residuals.

• Independence. We check this assumption by looking at the residual plot (upper right plot). This assumption also holds since there is not a specific clustering in the residuals.

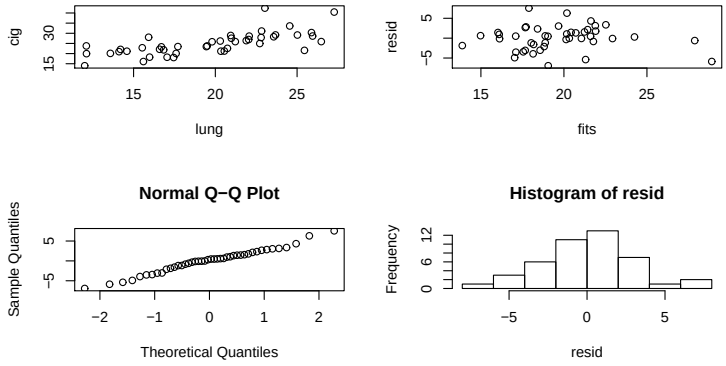


Figure 2: Residual plots for the simple linear regression of number of cigarettes on number of deaths..

Based on these plots, there are not any outliers in the residuals.

Question 4: Testing the significance of the linear relationship observed in the data.

(i) The P-value of the test is equal to $1.44e-07$ which is approximately equal to zero.

(ii) Since the p-value is very small we reject the null hypothesis, that is, the relationship between the two variables is significant.

(iii) We conclude that the number of deaths from lung cancer is significantly associated to the number of cigarettes smoked.

Question 5: Prediction. The prediction value for a for a number cigarettes to be smoked equal to 10 (hds per capita) is: $\widehat{y} = 6.4717 + 0.5291 * 10 = 11.7627$. The

prediction interval is derived from

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) \pm \widehat{t}_{\alpha/2,(n-2)}\sigma\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}.$$

For this we need to compute the average $\bar{x} = 24.914$ (you may use the *mean* function in R) and we need to obtain S_{xx}. Given the R summary of the regression, the easiest way to compute it is using

$$se(\beta_1)^2 = \frac{\widehat{\sigma}^2}{S_{xx}} \;\Rightarrow\; S_{xx} = \frac{\widehat{\sigma}^2}{se(\beta_1)^2} = \frac{3.066^2}{0.0839^2} = 1335.428.$$

Therefore, the confidence interval for the regression line at $x^* = 10$ is

$$11.7627 \pm t_{0.005,42}3.066\sqrt{1 + \frac{1}{44} + \frac{(10 - 24.914)^2}{1335.428}} = (2.74142520.783975).$$

The confidence interval is wide since the prediction is an extrapolation - the value at which we want to predict $x^* = 10$ is away from the center of the x values, $\bar{x} = 24.914$.