STAT7203: Applied Probability and Statistics
Quiz 11

1. Complete the practice exam.

   Solution: The solution to the practice exam is given elsewhere on Blackboard.

2. According to Hubble's law, relative velocity v (km/s) of any two galaxies separated by a distance D (Mega parsec – 1 parsec is 3.09×10^{13} km) is given by

$$v = H_0 D,$$

   where H_0 is Hubble's constant. If the expansion of the universe was linear, then $1/H_0$ (Hubble Time) would give the age of the universe. The velocities and distances of 24 galaxies containing Cepheid stars is given in the file hubble.xlsx.

   (a) Fit the linear regression model $V = \beta_0 + \beta_1 D + \varepsilon$, where $\varepsilon \sim \mathsf{N}(0, \sigma^2)$ to the hubble data. Assess the suitability of the linear regression model with diagnostic plots.

   Solution: With data stored in `hubble`, we fit the linear regression model using

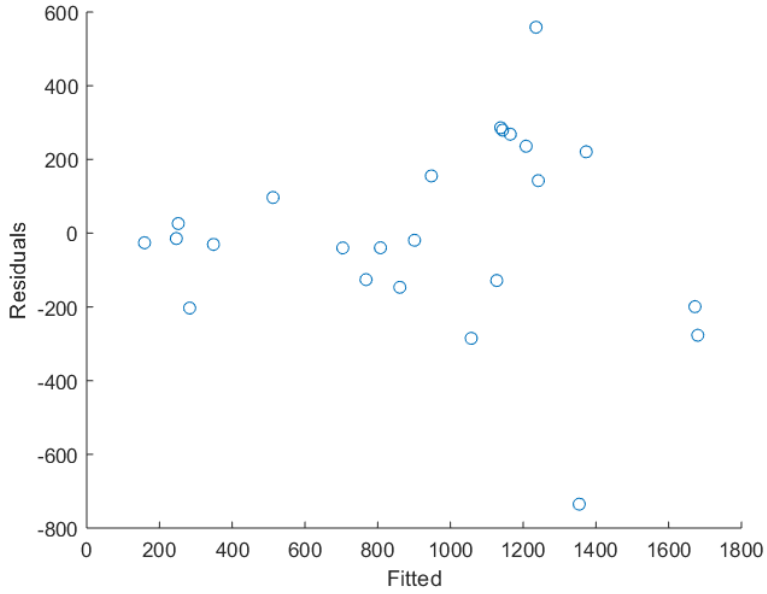
```
hubblelm = fitlm(hubble,'velocity distance')
```

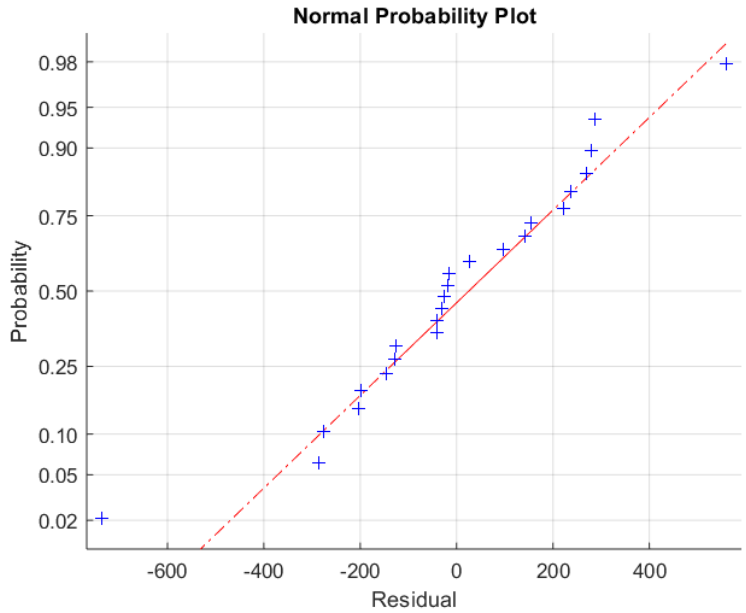
   The output is given below

```
Linear regression model:
    velocity ~ 1 + distance

Estimated Coefficients:
                   Estimate      SE       tStat       pValue
                   ________    ______    ________    _________

    (Intercept)      6.6963     126.56    0.052911      0.95828
    distance         76.127     9.4935      8.0189    5.6767e-08


Number of observations: 24, Error degrees of freedom: 22
Root Mean Squared Error: 265
R-squared: 0.745,  Adjusted R-Squared 0.733
F-statistic vs. constant model: 64.3, p-value = 5.68e-08
```

   To check the suitability of the linear regression model we plot the residuals versus fitted values and the normal probability plot. The plot of velocity against distance is given in the lecture notes. These plot can be generated using the code:

```
scatter(hubblelm.Fitted,hubblelm.Residuals.Raw)
xlabel('Fitted')
ylabel('Residuals')

normplot(hubblelm.Residuals.Raw)
xlabel('Residuals')
```





Neither plot indicates a clear departure from the model assumptions.

- The residuals appear randomly scatted about zero in the first plot. This is consistent with the mean response being linear in the explanatory variable.
- The variability of the residuals does not appear to change greatly with the fitted values which is consistent with the assumption of constant variance. This is difficult to detect with the rather small sample size. The large residual (approximately -800 around a fitted value of 1300) seems a little unusual. This is also reflected in the normal probability plot.

- The normal probability plot appears to be roughly a straightline which is consistent with the assumption of normality. Again the large negative residual (approximately -800) is the only indication of a possible departure from normality.

(b) Assuming the linear regression model is appropriate, is the data consistent with $\beta_0 = 0$.

Solution: The reported p-value for testing $H_0 : \beta_0 = 0$ against $H_1 : \beta_0 \neq 0$ is 0.958. (See the `Intercept` row in the output). There is no evidence against the null hypothesis, indicating that β_0 is zero.

(c) In the linear regression model, Hubble's constant is β_1. Construct a 99% confidence interval for the Hubble constant.

Solution: We can compute this in MATLAB using

```
coeffCI(hubblelm,0.01)

ans =

-350.0372  363.4297
   49.3672  102.8867
```

The first row is the confidence interval on the intercept term and the second row is the confidence interval on β_1. So we are 99% confident that Hubble's constant is between 49.3672 and 102.8867. *Try constructing the confidence interval just from the output of fitlm.*

(d) Construct a 95% confidence interval for the mean relative velocity of two galaxies separated by 10 Mega pasecs.

Solution: We can compute this in MATLAB using

```
[yhat, CI]=predict(hubblelm,[1 10],'Alpha',0.05)

yhat =

  767.9658


CI =

  648.8190  887.1126
```

So we are 95% confident that the mean relative velocity for two galaxies separated by 10 Mega parsecs is between 648.819 and 887.1126 (Km/s). *Try constructing the confidence interval using the output of fitlm and* `hubblelm.CoefficientCovariance`.

3. A study examined the relationship between SAT scores and GPAs of 105 students majoring in computer science at an American university. The data is available from `onlinestatbook.com/2/case_studies/sat.html`. Read the description of the dataset there and answer the following questions.

   (a) Fit the linear regression model to predict a student's overall university GPA from their mathematics and verbal scores in the SAT exam. Are the assumptions of the linear regression model satisfied?

   Solution: With data stored in `sat`, we fit the linear regression model using

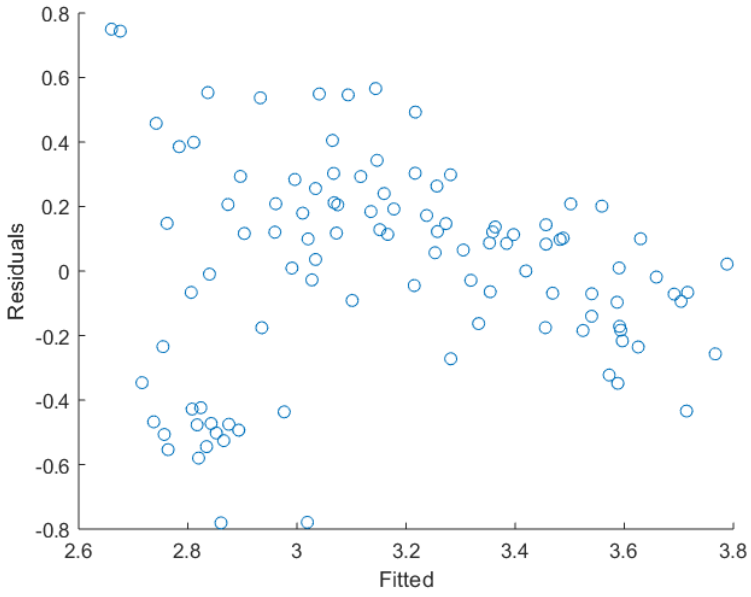
```
satlm = fitlm(sat,'univ_GPA ~ math_SAT + verb_SAT')
```

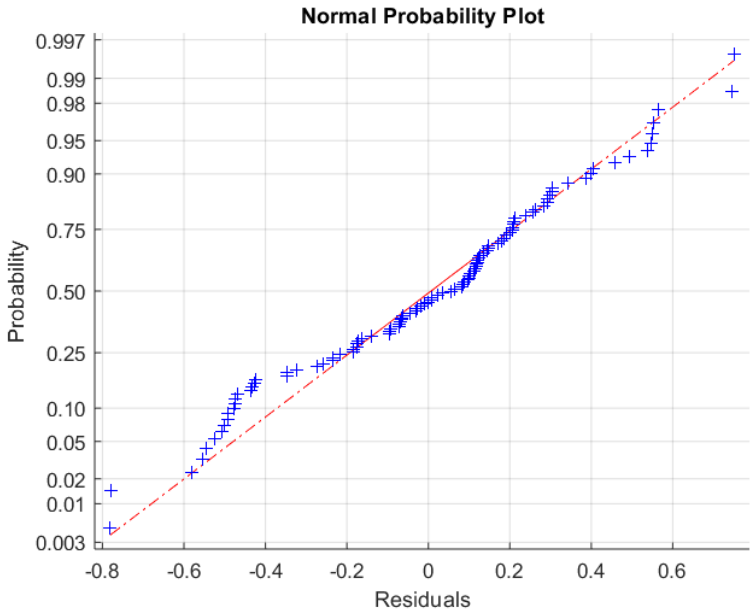
   The output is given below

```
Linear regression model:
    univ_GPA ~ 1 + math_SAT + verb_SAT

Estimated Coefficients:
                   Estimate        SE         tStat       pValue
                   ________    __________    ________    _________

    (Intercept)     -0.23753      0.37504    -0.63336      0.52792
    math_SAT       0.0032909    0.0010902      3.0187    0.0032068
    verb_SAT       0.0022718   0.00093082      2.4407     0.016383


Number of observations: 105, Error degrees of freedom: 102
Root Mean Squared Error: 0.329
R-squared: 0.47,  Adjusted R-Squared 0.46
F-statistic vs. constant model: 45.3, p-value = 8.49e-15
```

   To check the suitability of the linear regression model we plot the residuals versus fitted values and the normal probability plot. The plot of velocity against distance is given in the lecture notes. These plot can be generated using the code:

```
scatter(satlm.Fitted,satlm.Residuals.Raw)
xlabel('Fitted')
ylabel('Residuals')

normplot(satlm.Residuals.Raw)
xlabel('Residuals')
```





The normal probability plot looks ok, but there appears to be some structure in the residual vs fitted values plot. The spread is much greater for the smaller fitted values than larger fitted values. Furthermore, the residuals with fitted values greater than 3.4 tend to be negative while residuals with fitted values between 3 and 3.4 tend to be positive. There is potentially some other variable which affects the students GPA than we have not accounted for or there is a nonlinear relationship between the variables. For the remainder of this quiz we will suppose that the assumptions of the linear regression model are satisfied.

(b) Write out the fitted regression line and briefly interpret each of the estimated coefficients.

Solution: The fitted regression line is

$$\text{univGPA} = -0.2375 + 0.00329 \times \text{mathSAT} + 0.00227 \times \text{verbSAT} + \varepsilon.$$

- A person who scored 0 in the mathematics and verbal SAT exams has a

mean GPA of -0.2375.

- A unit increase in mathematics SAT score is associated with a 0.00329 unit increase in GPA.
- A unit increase in verbal SAT score is associated with a 0.00227 unit increase in GPA.

(c) Is there evidence of an association between overall GPA and mathematics SAT scores? What about an association between overall GPA and verbal SAT scores?

Solution: The p-value for the test of $H_0: \beta_{math} = 0$ against $H_1: \beta_{math} \neq 0$ is 0.0032. This is strong evidence against the null hypothesis, suggesting an association between mathematics SAT scores and university GPA. The p-value for the test of $H_0: \beta_{verb} = 0$ against $H_1: \beta_{verb} \neq 0$ is 0.0032. This is moderate evidence against the null hypothesis, suggesting an association between verbal SAT scores and university GPA.

(d) Give the covariance matrix of the estimator of the coefficients in the linear regression model.

Solution: We can compute this in MATLAB using

```
satlm.CoefficientCovariance

ans =

   0.140653363005865  -0.000233160603318   0.000009442454704
  -0.000233160603318   0.000001188458196  -0.000000847543271
   0.000009442454704  -0.000000847543271   0.000000866424287
```

You may need to set `format long`.

(e) Suppose a student scored 640 on mathematics SAT and 550 on verbal SAT. Provide a 99% confidence interval for their mean overall GPA.

Solution: We can compute this in MATLAB using

```
[yhat, CI]=predict(satlm,[1 640 550],'Alpha',0.01)

yhat =
3.118141

CI =
2.936096  3.300185
```

So we are 99% confident that the mean overall GPA of a student with 640 on mathematics SAT and 550 on verbal SAT is between 2.936096 and 3.300185. *Try constructing the confidence interval using the output of fitlm and* `satlm.CoefficientCovariance`.