Regression Inference and Forecasting – Solutions
STAT-UB.0103 – Statistics for Business Control and Regression Models

Inference

1. Here are the least squares estimates from the fit to model

$$\text{Price} = \beta_0 + \beta_1 \,\text{Size} + \varepsilon,$$

where price is measured in units of \$1000 and size is measured in units of 100 ft^2.

```
The regression equation is
price = 182 + 45.0 size

              Coef    SE Coef        T        P
Constant    182.27      62.43     2.92    0.010
size         44.95       4.37    10.29    0.000

S = 101.4    R-Sq = 86.9%    R-Sq(adj) = 86%
```

(a) Construct a 95% confidence interval for β_1.

> **Solution:** We use
>
> $$\hat{\beta}_1 \pm t_{\alpha/2}\text{SE}(\hat{\beta}_1),$$
>
> where $\alpha = .05$ and we have $n - 2 = 16$ degrees of freedom. This gives
>
> $$44.95 \pm 2.120(4.37) = 44.95 \pm 9.26,$$
>
> or $(35.69, 54.21)$.

(b) What is the meaning of the confidence interval for β_1?

> **Solution:** We are 95% confident that if we increase size by 100 square feet, then mean price will increase by an amount between \$35.7K adn \$54.2K.

(c) What is the meaning of a 95% confidence interval for β_0? In the context of the housing data, is this useful?

> **Solution:** This would be a confidence interval for the mean price of apartments with size 0. This is nonsensical (no apartments have size 0), and thus not useful.

(d) Perform a hypothesis test at level 5% of whether or not the is a linear relationship between Size and mean Price.

> **Solution:** We are interested in the following null and alternative hypotheses:
>
> $$H_0 : \beta_1 = 0 \quad \text{(no linear relationship)}$$
> $$H_a : \beta_1 \neq 0 \quad \text{(linear relationship)}$$

Based on the minitab output, the p-value for this test is below 0.001. Thus, we reject the null hypothesis at level 5%. There is a statistically significant linear relationship between size and mean price.

We can also do this problem using a rejection region. We reject H_0 at level α if $|T| > t_{\alpha/2}$, where

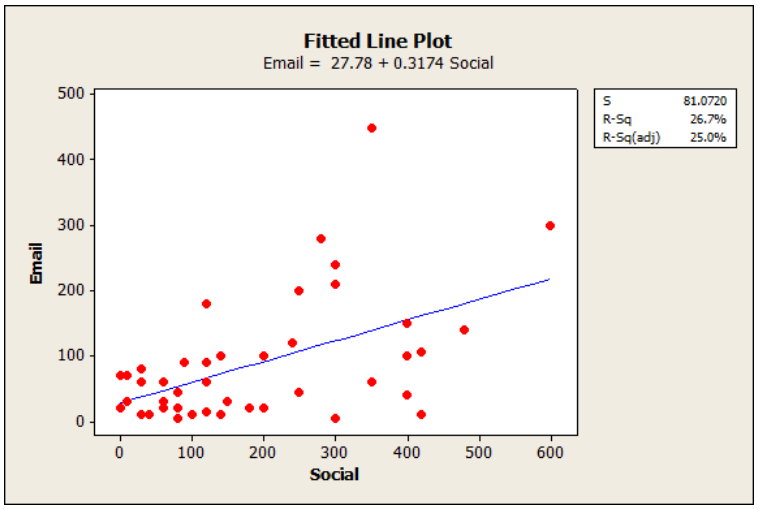
$$T = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} = \frac{44.95}{4.37} = 10.286$$

For level $\alpha = .05$, we have $t_{\alpha/2} = t_{.025} = 2.120$ (using $n - 2 = 16$ degrees of freedom). Since $|T| > 2.120$, we reject H_0.

2. 44 NYU undergraduates reported the amount of time they spent communicating via email and via social media (in minutes per week). We will use this data to examine the relationship between email usage and social media usage. We fit the model

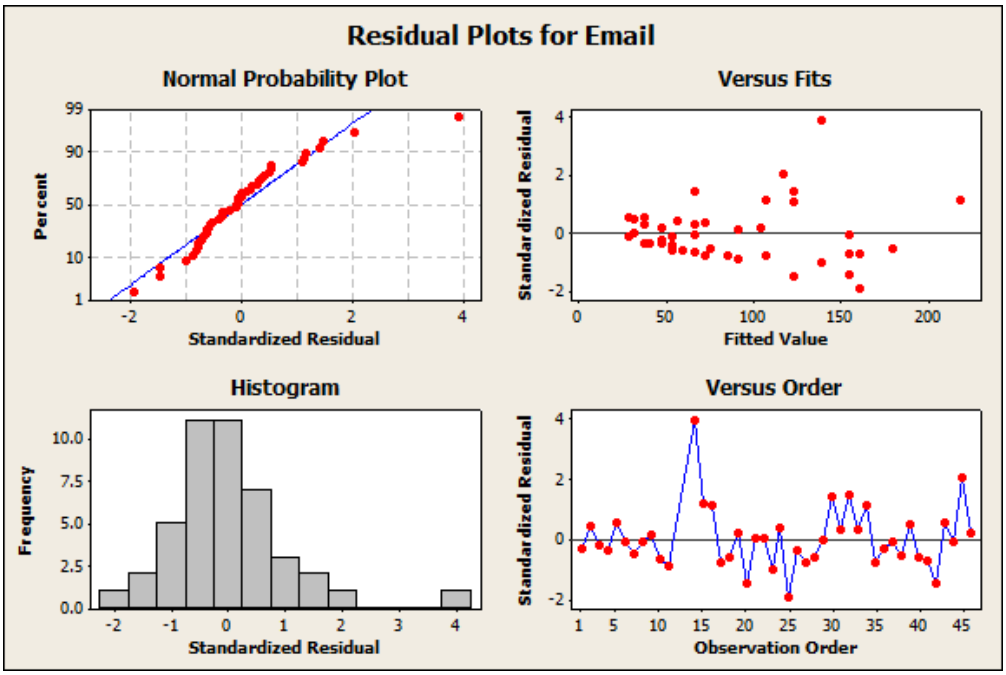
$$\text{Email} = \beta_0 + \beta_1 \text{Social} + \varepsilon.$$

using least-squares. The scatterplot and Minitab regression output follow.




```
The regression equation is
Email = 27.8 + 0.317 Social

Predictor      Coef  SE Coef     T      P
Constant      27.78    19.32  1.44  0.158
Social      0.31744  0.08109  3.91  0.000

S = 81.0720   R-Sq = 26.7%   R-Sq(adj) = 25.0%


Analysis of Variance

Source           DF      SS      MS      F      P
Regression        1  100716  100716  15.32  0.000
Residual Error   42  276052    6573
Total            43  376768


Unusual Observations

Obs  Social  Email    Fit  SE Fit  Residual  St Resid
 14     350  450.0  138.9    18.1     311.1      3.94R
 16     600  300.0  218.2    35.8      81.8      1.12 X
 45     280  280.0  116.7    14.5     163.3      2.05R
```

(a) Use the residual plots below to assess whether or not the regression assumptions hold.



Solution: There is one large residual which is evidence of non-normality. Also, there may be some non-constant variance (the variance of the residuals appears to increase with $\hat{y}_i$). However, overall, the assumptions look reasonable.

(b) Is there a significant linear relationship between social media usage and email usage?

Solution: Yes, the p-values is below 0.001 (it is reported as 0.000).

(c) Quantify the relationship between email usage and social media usage using a 95% confidence interval. (You will need the value $t_{.025} = 2.018$.)

Solution: We do this by finding a confidence interval for β_1. We use

$$\hat{\beta}_1 \pm t_{\alpha/2}\text{SE}(\hat{\beta}_1),$$

where $\alpha = .05$ and we have $n - 2 = 44$ degrees of freedom. This gives

$$0.31744 \pm 2.018(0.08109) = 0.31744 \pm 0.16364,$$

or $(0.15380, 0.48108)$. When we interpret this confidence interval, we should take not of the points we removed. For example, one statement we can make with 95% confidence is that when social media usage is below 600 minutes per week, for every additional 10 minutes per week spent communicating via social media, the mean amount sent communicating via email increases by an amount between 1.5 and 4.8 minutes per week.

Forecasting

3. Here are the least squares estimates from the fit to model Price $= \beta_0 + \beta_1$ Size $+ \varepsilon$, where price is measured in units of \$1000 and size is measured in units of 100 ft^2, along with the result of using the model to predict the mean price at size 2000 ft^2.

```
The regression equation is
price = 182 + 45.0 size

              Coef    SE Coef        T        P
Constant    182.27      62.43     2.92    0.010
size         44.95       4.37    10.29    0.000

S = 101.4    R-Sq = 86.9%    R-Sq(adj) = 86%

Predicted Values for New Observations

NewObs     Fit  SE Fit       95% CI            95% PI
      1  1081.3    38.1  (1000.4, 1162.1)  (851.7, 1310.9)

Values of Predictors for New Observations

NewObs           size
      1           20.0
```

(a) Find a 95% confidence interval for the mean price of all apartments with size 2000 ft^2.

> **Solution:** This is given in the output: (1000.4, 1162.1). We 95% confidence, the mean price of all apartments with size 2000 ft^2 is between \$1,000,400 and \$1,152,100.

(b) Find a 95% prediction interval for the price of a particular apartments with size 2000 ft^2.

> **Solution:** Again, this is given in the output: (851.7, 1310.9). If someone tells us that a particular apartment has size 2000 ft^2, then we can say with 95% confidence that the price of the apartment is between \$851,700 and \$1,310,900.

(c) Make a statement about the prices of 95% of all apartments with size 2000 ft^2.

> **Solution:** To make a statement about *all* apartments, we use a prediction interval. With 95% confidence, 95% of all apartments with size 2000 ft^2 have sizes between \$851,700 and \$1,310,900.

(d) What is the difference between the confidence interval and the prediction interval?

Solution: A confidence interval is a statement about the mean value of Y; a prediction interval is a statement about a particular value of Y (equivalently, all values of Y).