**Multiple Regression, $F$ Tests (Solutions)**
STAT-UB.0003: Regression and Forecasting Models

# Multiple Regression

1. We used $n = 294$ from the 2003 Zagat restaurant guide for New York City to fit a regression model, with "Price" as the response variable and "Food," "Decor," and "Service" as predictor variables. Here is the output:

```
Analysis of Variance

Source          DF   Adj SS   Adj MS  F-Value  P-Value
Regression       3  49418.0  16472.7   330.49    0.000
  Food           1     19.1     19.1     0.38    0.537
  Decor          1   3257.8   3257.8    65.36    0.000
  Service        1   5938.5   5938.5   119.14    0.000
Error          290  14454.5     49.8
  Lack-of-Fit  245  12075.7     49.3     0.93    0.640
  Pure Error    45   2378.8     52.9
Total          293  63872.5

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
7.05997  77.37%     77.14%      76.68%

Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant  -20.69     2.31    -8.96    0.000
Food      -0.103    0.167    -0.62    0.537  2.21
Decor      1.026    0.127     8.08    0.000  2.33
Service    2.555    0.234    10.92    0.000  4.05

Regression Equation

Price = -20.69 - 0.103 Food + 1.026 Decor + 2.555 Service
```

   (a) Interpret the coefficient of "Food" in the context of the estimated multiple regression model. How can this value be negative?

   **Solution:** In a regression model with Food, Decor, and Service, increasing Food by 1 point while holding all other predictors constant decreases the mean value of Price by 0.10.

   This is saying that when comparing restaurants with the same Decor and Service, those with higher Food quality tend to be cheaper on average.

   (b) Does "Food" have utility in explaining "Price" beyond what is explained by "Decor" and "Service"?

**Solution:** To answer this question, we perform a test with the hypotheses

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

The p-value is given in the minitab output as $p = .537$. Thus, there is no significant evidence (at level .05) that Food has utility in explaining Price usage beyond what is explained by Decor and Service.

(c) Give a 95% confidence interval for the amount that mean price goes up when we increase food quality rating by 1 point but we hold decor and service ratings constant.

**Solution:** With $\alpha = .05$ and $n - k - 1 = 294 - 3 - 1 = 290$ degrees of freedom, we have $t_{\alpha/2} \approx z_{.025} \approx 2$. The 95% confidence interval for $\beta_1$ is

$$\hat{\beta}_1 \pm t_{\alpha/2} \cdot \mathrm{SE}(\hat{\beta}_1),$$
$$-0.1034 \pm 2 \cdot 0.1672,$$
$$-0.1034 \pm 0.3344,$$

or $(-0.7722, 0.5654)$.

2. In the previous problem, we found that "Food" was not useful for explaining "Price" after adjusting for "Decor" and "Service." After removing "Food" from the regression model, we get a new regression fit:

```
Analysis of Variance

Source          DF  Adj SS   Adj MS  F-Value  P-Value
Regression       2   49399  24699.5   496.60    0.000
  Decor          1    3802   3802.2    76.45    0.000
  Service        1   10586  10586.2   212.84    0.000
Error          291   14474     49.7
  Lack-of-Fit  143    7232     50.6     1.03    0.421
  Pure Error   148    7241     48.9
Total          293   63873

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
7.05247  77.34%     77.18%      76.84%

Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant  -21.39     2.01   -10.63    0.000
Decor      1.051    0.120     8.74    0.000  2.10
Service    2.455    0.168    14.59    0.000  2.10

Regression Equation

Price = -21.39 + 1.051 Decor + 2.455 Service
```

Use this regression model to answer the following questions.

(a) Interpret the coefficient of "Service" in the context of the estimated multiple regression model.

> **Solution:** In a regression model with Decor and Service, increasing Service by 1 point while holding Decor constant increases the mean value of Price by \$2.45.

(b) Does Service have utility in explaining Price beyond what is explained by Decor?

> **Solution:** To answer this question, we perform a test with the hypotheses
>
> $$H_0 : \beta_2 = 0$$
> $$H_a : \beta_2 \neq 0$$
>
> The p-value is given in the minitab output as $p = 0.000$. Thus, there is significant evidence (at level 0.1%) that Service has utility in explaining Price beyond what is explained by Decor.

(c) Give a 95% confidence interval for the amount that mean Price goes up when we increase Service by 1 point but we hold Decor constant.

**Solution:** With $\alpha = .05$ and $n - k - 1 = 294 - 2 - 1 = 291$ degrees of freedom, we have $t_{\alpha/2} \approx z_{.025} = 2$. The 95% confidence interval for $\beta_2$ is

$$\begin{aligned} \hat{\beta}_2 &\pm t_{\alpha/2} \cdot \mathrm{SE}(\hat{\beta}_2), \\ 2.4546 &\pm 2 \cdot 0.1682, \\ 2.4546 &\pm 0.3364 \end{aligned}$$

or $(2.1182, 2.7910)$.

## Regression $F$ Tests

3. Locate the regression $F$ statistic and the corresponding $p$ value in the output from the previous problem.

   (a) How is the regression $F$ statistic computed?

   **Solution:**
   $$F = \frac{MSR}{MSE} = \frac{38186}{6714} = 5.69.$$

   (b) How many numerator and denominator degrees of freedom are there in the regression $F$ statistic?

   **Solution:** $k = 3$ numerator degrees of freedom; $n - k - 1 = 39$ denominator degrees of freedom.

   (c) How is the $p$-value computed?

   **Solution:** We find $P(F \geqslant 5.69)$, the probability that an F-distributed random variable with 3 numerator degrees of freedom and 39 denominator degrees of freedom is greater than or equal to 5.69. This can be done using an $F$ table, or by using Minitab. (You are not expected to know how to use an $F$ table.)

   (d) What are the null and alternative hypothesis for the regression $F$ test?

   **Solution:**

   $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ (the regression model is useless)
   $H_1 : \beta_j \neq 0$ for some $j = 1, 2,$ or $3$ (the regression model has use in explaining email)

   (e) Based on the $p$-value, what is the conclusion of the regression $F$ test (use a significance level of 5%)?

   **Solution:** The $p$-value is 0.002, which is less than $\alpha = .05$. Thus, we reject the null hypothesis at level 5%. There is evidence that the model is useful for explaining email usage.

## More Multiple Regression

4. We have a dataset measuring the price ($), size ($ft^2$), number of bedrooms, and age (years) of 518 houses in Easton, Pennsylvania. We fit a regression model to explain price in terms of the other variables.

```
Analysis of Variance

Source          DF       Adj SS       Adj MS  F-Value  P-Value
Regression       3  85029785549  28343261850   178.18    0.000
  SIZE           1  53484452975  53484452975   336.24    0.000
  BEDROOM        1    156773465    156773465     0.99    0.321
  AGE            1    279354141    279354141     1.76    0.186
Error          514  81760176401    159066491
  Lack-of-Fit  509  80933266401    159004453     0.96    0.607
  Pure Error     5    826910000    165382000
Total          517  1.66790E+11

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
12612.2  50.98%     50.69%      50.19%

Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant  25875     3555     7.28    0.000
SIZE      39.20     2.14    18.34    0.000  1.71
BEDROOM   -1145     1153    -0.99    0.321  1.71
AGE        -354      267    -1.33    0.186  1.01

Regression Equation

PRICE = 25875 + 39.20 SIZE - 1145 BEDROOM - 354 AGE
```

(a) Do the signs of the coefficients make sense to you? Explain any apparent contradictions between what you would expect and what the Minitab output indicates.

**Solution:**
We would expect Price to be positively associated with Size and Bedroom (bigger houses tend to be more expensive), but negatively associated with Age (older houses tend to be cheaper). However, in the multiple regression model with all three variables as predictors, the coefficient of Bedroom is negative. We can explain this apparent contradiction by noting that the regression coefficient measures the change in mean price when Bedroom is increased *and all other predictors are held constant*. If we hold Size constant while increasing Bedroom, then the bedrooms get smaller.

(b) What does the result of the $t$ test on the coefficient of Size indicate?

> **Solution:** The coefficient is significant ($p < 0.001$). Size has the ability to explain Price beyond what is explained by Bedroom and Age.

(c) What does the result of the $t$ test on the coefficient of Bedroom indicate?

> **Solution:** The coefficient is not significant ($p = 0.321$). Bedroom does not convey additional information in explaining Price Price beyond what is explained by Size and Age.

(d) What does the result of the regression $F$ test indicate?

> **Solution:** The test statistic is significant ($p < 0.001$). Thus, there is statistically significant evidence that the model is useful in explaining Price.