

Regression Inference and Forecasting – Solutions
COR1-GB.1305 – Statistics and Data Analysis

Inference

1. Here are the least squares estimates from the fitting the model

$$\text{Price} = \beta_0 + \beta_1 \text{Size} + \varepsilon_1$$

for $n = 18$ apartments in Greenwich Village. Price is measured in units of \$1000 and size is measured in units of 100 ft².

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
101.375	86.87%	86.05%	81.13%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	182.3	62.4	2.92	0.010	
Size(100sqft)	44.95	4.37	10.29	0.000	1.00

Regression Equation

$$\text{Price}(\$1000) = 182.3 + 44.95 \text{ Size}(100\text{sqft})$$

- (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 and \$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 there 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

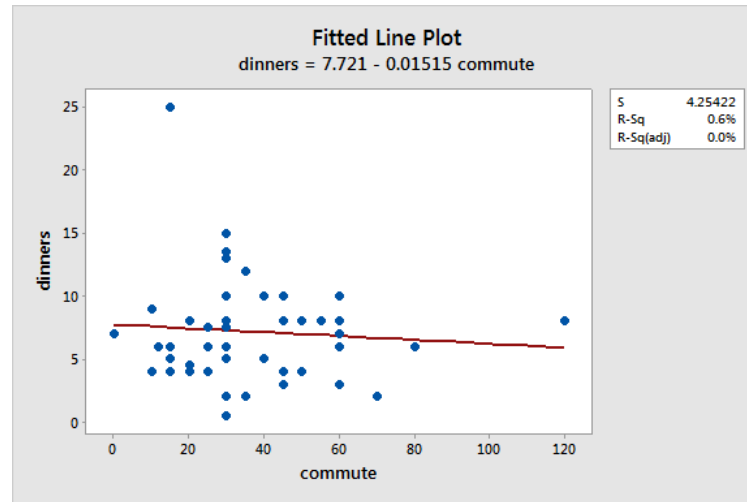
$$T = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} = \frac{44.95}{4.37} = 10.286$$

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. 51 students reported their commute times (in minutes) and the number of times they go out to dinner in a typical month. We will use this data to examine the relationship between these two variables. We fit the model

$$\text{Dinners} = \beta_0 + \beta_1 \text{Commute} + \varepsilon$$

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



Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4.25422	0.56%	0.00%	0.00%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	7.72	1.21	6.38	0.000	
commute	-0.0152	0.0288	-0.53	0.601	1.00

Regression Equation

$\text{dinners} = 7.72 - 0.0152 \text{ commute}$

- (a) Quantify the relationship between dinners and commute time using a 95% confidence interval. (You will need the value $t_{.025,49} \approx 2.021$.)

Solution: A 95% confidence interval for β_1 is

$$\begin{aligned} \hat{\beta}_1 \pm t_{.025, n-2} \text{se}(\hat{\beta}_1) &= (-0.0152) \pm (2.021)(0.0288) \\ &= -0.0152 \pm 0.0582 \\ &= (-0.0734, 0.0430) \end{aligned}$$

We can be 95% confident that increasing commute time by 1 minute is associated with changing expected number of dinners eaten out per month by an amount between -0.0734 and 0.0430 .

- (b) Perform a hypothesis test to determine if there is a significant linear relationship between dinners and commute time.

Solution: The null and alternative hypotheses are

$$H_0 : \beta_1 = 0 \quad (\text{no linear relationship})$$

$$H_a : \beta_1 \neq 0 \quad (\text{linear relationship})$$

The p -value for this test is $p = 0.601$. Since $p \geq 0.05$, we do not reject H_0 ; there is no significant relationship between dinners and commute time.

Forecasting

3. We used the regression model fit to the housing data to predict price at size 2000 ft²:

Regression Equation

$$\text{Price}(\$1000) = 182.3 + 44.95 \text{ Size}(100\text{sqft})$$

Variable	Setting
Size(100sqft)	20

Fit	SE Fit	95% CI	95% PI
1081.27	38.1287	(1000.44, 1162.10)	(851.667, 1310.88)

- (a) Find a 95% confidence interval for the mean price of all apartments with size 2000 ft².

Solution: This is given in the output: (1000.4, 1162.1). We 95% confidence, the mean price of all apartments with size 2000 ft² 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 ft².

Solution: Again, this is given in the output: (851.7, 1310.9). If someone tells us that a particular apartment has size 2000 ft², 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 ft².

Solution: To make a statement about *all* apartments, we use a prediction interval. With 95% confidence, 95% of all apartments with size 2000 ft² 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).

4. We fit a regression model to the 294 restaurants from the 2003 Zagat data. Our predictor variable is food quality (1–30), and our response variable is price (\$). Here is the result of using the fitted model to predict the price when the food quality is 25.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
12.5559	27.93%	27.68%	26.86%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-4.74	3.95	-1.20	0.232	
Food	2.129	0.200	10.64	0.000	1.00

Regression Equation

Price = -4.74 + 2.129 Food

Variable	Setting
Food	25

Fit	SE Fit	95% CI	95% PI
48.4832	1.33906	(45.8478, 51.1187)	(23.6315, 73.3349)

- (a) What is the interpretation of the 95% confidence interval?

Solution: We are 95% confident that the average price of all 2003 New York City restaurants with quality ratings of 25 is between \$45.84 and \$51.12.

- (b) What is the interpretation of the 95% prediction interval?

Solution: Approximately 95% of all 2003 New York City restaurants with quality ratings of 25 have prices between \$23.63 and \$73.34.

- (c) Explain how the confidence interval is related to Fit, SE Fit, and S.

Solution: The 95% confidence interval for $E(Y \mid x = 25)$ is approximately equal to

$$\hat{y}(25) \pm 2se(\hat{y}(25)) = 48.4832 \pm (2)(1.33906).$$

(For an exact equivalence, use $t_{.025, n-2}$ instead of 2.)

- (d) Explain how the prediction interval is related to Fit, SE Fit, and S.

Solution: The 95% prediction interval is approximately equal to

$$\hat{y}(25) \pm 2s = 48.4832 \pm (2)(12.5559).$$

(For an exact equivalence you would use the formula

$$\hat{y}(x) \pm t_{.025, n-2} \sqrt{s^2 + [\text{se}\{\hat{y}(x)\}]^2};$$

you are not expected to know this formula.)