HABIB UNIVERSITY

Final Exam (max pt. 100)
Dr. Shafayat Abrar
CE 362: Statistics and Inferencing

Timing: 02:30 - 05:30 PM
Dated: Dec. 15, 2023
Duration: 120 min

Material 01:

Confidence Intervals on Parameters

It is possible to obtain confidence interval estimates of parameters of linear model. The width of these confidence intervals is a measure of the overall quality of the regression line. If the error terms, ϵ_i, in the regression model are normally and independently distributed,

$$\frac{\widehat{\beta}_1-\beta_1}{\sqrt{\dfrac{\widehat{\sigma}^2}{S_{xx}}}} \quad \text{and} \quad \frac{\widehat{\beta}_0-\beta_0}{\sqrt{\widehat{\sigma}^2\left[\dfrac{1}{n}+\dfrac{\overline{x}^2}{S_{xx}}\right]}}$$

are both distributed as t random variables with $n-2$ degrees of freedom. Under the assumption that the observations are normally and independently distributed, a $100(1-\alpha)\%$ confidence interval on the slope β_1 in simple linear regression is

$$\widehat{\beta}_1-t_{\alpha/2,n-2}\sqrt{\frac{\widehat{\sigma}^2}{S_{xx}}}\leq\beta_1\leq\widehat{\beta}_1+t_{\alpha/2,n-2}\sqrt{\frac{\widehat{\sigma}^2}{S_{xx}}}$$

Similarly, a $100(1-\alpha)\%$ confidence interval on the intercept β_0 is

$$\widehat{\beta}_0-t_{\alpha/2,n-2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right]}\leq\beta_0\leq\widehat{\beta}_0+t_{\alpha/2,n-2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right]}$$

Confidence Intervals on the Mean Response

A confidence interval may be constructed on the mean response (of y) at a specified value of x, say, x_0. This is a confidence interval about $E\left[Y\mid x_0\right]=:\mu_{Y|x_0}$ and is sometimes referred to as a confidence interval about the regression line. Because $E\left[Y\mid x_0\right]=\mu_{Y|x_0}=\beta_0+\beta_1x_0$, we may obtain a point estimate of the mean of Y at $x=x_0$ from the fitted model as

$$\widehat{\mu}_{Y|x_0}=\widehat{\beta}_0+\widehat{\beta}_1x_0$$

Now $\widehat{\mu}_{Y|x_0}$ is an unbiased point estimator of $\mu_{Y|x_0}$ because $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased estimators of β_0 and β_1. The variance of $\widehat{\mu}_{Y|x_0}$ is

$$\operatorname{var}\left(\widehat{\mu}_{Y|x_0}\right)=\sigma^2\left[\frac{1}{n}+\frac{\left(x_0-\overline{x}\right)^2}{S_{xx}}\right]$$

This last result follows from the fact that $\widehat{\mu}_{Y|x_0}=\overline{y}+\widehat{\beta}_1\left(x_0-\overline{x}\right)$ and $\operatorname{cov}\left(\overline{Y},\widehat{\beta}_1\right)=0$. The zero covariance result is left as a mind-expanding exercise. Also, $\widehat{\mu}_{Y|x_0}$ is normally distributed because $\widehat{\beta}_1$

and $\widehat{\beta}_0$ are normally distributed, and if we use $\widehat{\sigma}^2$ as an estimate of σ^2, it is easy to show that

$$\frac{\widehat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}}$$

has a t distribution with $n-2$ degrees of freedom. This leads to the following confidence interval definition. A $100(1-\alpha)\%$ confidence interval on the mean response of y at the value of $x = x_0$, say $\mu_{Y|x_0}$, is given by

$$\widehat{\mu}_{Y|x_0} - t_{\alpha/2,n-2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]} \le \mu_{Y|x_0} \le \widehat{\mu}_{Y|x_0} + t_{\alpha/2,n-2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}$$

where $\widehat{\mu}_{Y|x_0} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$ is computed from the fitted regression model. Note that the width of the confidence interval for $\mu_{Y|x_0}$ is a function of the value specified for x_0. The interval width is a minimum for $x_0 = \overline{x}$ and widens as $|x_0 - \overline{x}|$ increases.

Prediction Interval on a Future Observation

An important application of a regression model is predicting new or future observations Y corresponding to a specified level of the regressor variable x. If x_0 is a variable of interest, not available in the data-set (x, Y), then

$$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$$

is the point estimator of the new or future value of the response Y_0. Note that the error in prediction

$$e_p = Y_0 - \widehat{Y}_0$$

is a normally distributed random variable with mean zero and variance

$$\operatorname{var}(e_p) = \operatorname{var}\left(Y_0 - \widehat{Y}_0\right) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]$$

because Y_0 is independent of $\widehat{Y}_0$. If we use $\widehat{\sigma}^2$ to estimate σ^2, we can show that

$$\frac{Y_0 - \widehat{Y}_0}{\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}}$$

has a t distribution with $n-2$ degrees of freedom. From this, we can develop the following prediction interval definition. A $100(1-\alpha)\%$ prediction interval on a future observation Y_0 at the value x_0 is given by

$$\widehat{Y}_0 - t_{\alpha/2,n-2}\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]} \le Y_0 \le \widehat{Y}_0 + t_{\alpha/2,n-2}\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}$$

The value $\widehat{y}_0$ is computed from the regression model $\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$.

Q1 [30 points]: Refer to the data-set for **Q1** on Y = house selling price and x = taxes paid.

(a) Obtain a linear fit on the data.

(b) Find a 95% confidence interval on β_0.

(c) Find a 95% confidence interval on β_1.

(d) Find a 95% confidence interval on mean selling price when the taxes paid is $x = 8.0$.

(e) Compute the 95% prediction interval for selling price when the taxes paid is $x = 8.0$.

(f) Why prediction interval for selling price is far wider than confidence interval of mean selling price at the same value of x? explain briefly.

Material 02:

Regression on Transformed Variables

We occasionally find that the straight-line regression model $Y = \beta_0 + \beta_1 x + \epsilon$ is inappropriate because the true regression function is nonlinear. Sometimes nonlinearity is visually determined from the scatter diagram, and sometimes, because of prior experience or underlying theory, we know in advance that the model is nonlinear. Occasionally, a scatter diagram will exhibit an apparent nonlinear relationship between Y and x. In some of these situations, a nonlinear function can be expressed as a straight line by using a suitable transformation. Such nonlinear models are called **intrinsically linear**.

As an example of a nonlinear model that is intrinsically linear, consider the exponential function

$$Y = \beta_0 e^{\beta_1 x}\epsilon$$

This function is intrinsically linear because it can be transformed to a straight line by a logarithmic transformation

$$\ln Y = \ln \beta_0 + \beta_1 x + \ln \epsilon$$

This transformation requires that the transformed error terms $\ln \in$ are normally and independently distributed with mean 0 and variance σ^2.

Another intrinsically linear function is

$$Y = \beta_0 + \beta_1 \left(\frac{1}{x}\right) + \epsilon$$

By using the reciprocal transformation $z = 1/x$, the model is linearized to

$$Y = \beta_0 + \beta_1 z + \epsilon$$

Sometimes several transformations can be employed jointly to linearize a function. For example, consider the function

$$Y = \frac{1}{\exp\left(\beta_0 + \beta_1 x + \epsilon\right)}$$

letting $Y^* = 1/Y$, we have the linearized form

$$\ln Y^* = \beta_0 + \beta_1 x + \epsilon$$

Q2 [40 points]: A research engineer is investigating the use of a windmill to generate electricity and has collected data on the DC output from this windmill and the corresponding wind velocity. Data is available in Excel file

(a) Obtain a cluster plot of Y versus x. Sketch the plot in your answer sheet.

(b) Use regression to obtain a quadratic model such as

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

(c) Obtain a cluster plot and/or histogram of the estimation error $e = \widehat{Y} - Y$ for the quadratic model. Is this error zero-mean and Gaussian in nature?

(d) Now consider an intrinsically linear model of the form:

$$Y = \alpha_0 + \alpha_1 \frac{1}{x} + \epsilon = \alpha_0 + \alpha_1 z + \epsilon$$

where $z = \frac{1}{x}$ is an intrinsically linear variable. Obtain a cluster plot of Y versus z. What do you observe?

(e) Apply regression to obtain the values of α_0 and α_1.

(f) Obtain a cluster plot and/or histogram of the estimation error $e = \widehat{Y} - Y$ for intrinsically linear model. Is this error zero-mean and Gaussian in nature?

(g) In order to validate these models, the engineer collects three more data values (refer to Excel file). Find the estimation errors $e = \widehat{Y} - Y$ for both quadratic and intrinsically linear models using validation data. Take the sum of absolute errors for both models. What do you observe, which model is better in terms of sum of absolute errors? write down your observations.

Q3 [10 points]: Suppose that we have assumed the single-variable *simple linear* regression model $Y = \beta_0 + \beta_1 x_1 + \epsilon$, and the value of β_1 is estimated by using data x_1 as follows:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_{1.i} - \overline{x}_1) Y_i}{S^2_{x_1}}$$

But in reality, the true response Y is affected by a second variable x_2 such that the *true* regression function is $E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. The estimator $\widehat{\beta}_1$ of the slope β_1 in the simple linear regression model is no more ~~biased!~~ unbiased what is the value of the bias?

Q4 [20 points]: Safety in motels and hotels is a growing concern among travelers. Suppose a survey was conducted by the National Motel and Hotel Association to determine U.S. travelers' perception of safety in various motel chains. The association chose four different national chains from the economy lodging sector and randomly selected 10 people who had stayed overnight in a motel in each of the four chains in the past two years. Each selected traveler was asked to rate each motel chain on a scale from 0 to 100 to indicate how safe he or she felt at that motel. A score of 0 indicates completely unsafe and a score of 100 indicates perfectly safe. The scores follow. Test this randomized block design to determine whether there is a significant difference in the safety ratings of the four motels. Use $\alpha = 0.05$. Refer to Excel file for data.