Regression

Covariance and Correlation

Bruce M. Boghosian

Tufts University School of Arts and Sciences

Department of Mathematics

Tufts University

Outline

A theorem from probability

- Let $Y_1, \ldots, Y_n$ be any set of independent random variables with means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$, respectively. Let $a_1, \ldots, a_n$ be any set of constants. Then $Y = a_1 Y_1 + \cdots + a_n Y_n$ is normally distributed with mean $\mu = \sum_i^n a_i \mu_i$ and variance $\sigma^2 = \sum_i^n a_i^2 \sigma_i^2$.
- Proof is straightforward using moment-generating functions.

The Linear Model

- We are given data $(x_1, Y_1), \ldots, (x_n, Y_n)$.
- Linear Model supposes that $f_{Y|x}(y)$ is normal for all x.
- Standard deviation σ assumed independent of x.
- Means are collinear, $E(Y|x) = \beta_0 + \beta_1 x$.
- Results of maximum likelihood estimation

$$\hat{\beta}_0 = \frac{\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n Y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i Y_i\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2}$$

$$\hat{\beta}_1 = \frac{n\left(\sum_{i=1}^n x_i Y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n Y_i\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2}$$

- Same form obtained from geometric approach to fitting.

Theorems about Linear Model estimators

- **Thm.:** $\hat{\beta}_0$ and $\hat{\beta}_1$ are both normally distributed.
- **Thm.:** $\hat{\beta}_0$ and $\hat{\beta}_1$ are both unbiased,

$$E\left(\hat{\beta}_0\right) = \beta_0$$
$$E\left(\hat{\beta}_1\right) = \beta_1$$

- **Thm.:** The variances of the estimators are

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum_i^n x_i^2}{n \sum_i^n (x_i - \overline{x})^2}$$
$$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{n \sum_i^n (x_i - \overline{x})^2}$$

Proof of normal distribution

- **Pf. (for $\hat{\beta}_1$):** Note that

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{n\sum_i^n x_i Y_i - \left(\sum_i^n x_i\right)\left(\sum_i^n Y_i\right)}{n\left(\sum_i^n x_i^2\right) - \left(\sum_i^n x_i\right)^2} = \frac{\sum_i^n x_i Y_i - \left(\frac{1}{n}\sum_i^n x_i\right)\left(\sum_i^n Y_i\right)}{\left(\sum_i^n x_i^2\right) - n\left(\frac{1}{n}\sum_i^n x_i\right)\left(\frac{1}{n}\sum_i^n x_i\right)} \\
&= \frac{\sum_i^n x_i Y_i - \overline{x}\left(\sum_i^n Y_i\right)}{\left(\sum_i^n x_i^2\right) - n\overline{x}^2} = \frac{\sum_i^n \left(x_i - \overline{x}\right) Y_i}{\left(\sum_i^n x_i^2\right) - n\overline{x}^2}
\end{aligned}
$$

- This is a linear combination of normally distributed r.v.s, and thus normally distributed.

Proof of unbiasedness

- **Pf. (for $\hat{\beta}_1$):** Using the same form for $\hat{\beta}_1$ used above,

$$
\begin{aligned}
E\left(\hat{\beta}_1\right) &= E\left(\frac{\sum_i^n (x_i - \overline{x})\, Y_i}{\left(\sum_i^n x_i^2\right) - n\overline{x}^2}\right) \\
&= \frac{\sum_i^n (x_i - \overline{x})\, E\left(Y_i\right)}{\left(\sum_i^n x_i^2\right) - n\overline{x}^2} \\
&= \frac{\sum_i^n (x_i - \overline{x})\left(\beta_0 + \beta_1 x_i\right)}{\left(\sum_i^n x_i^2\right) - n\overline{x}^2} \\
&= \beta_1 .
\end{aligned}
$$

Proof of expressions for variances

- **Pf. (for $\hat{\beta}_1$):** Using the same form for $\hat{\beta}_1$ used above,

$$\begin{aligned}
\mathrm{Var}\left(\hat{\beta}_1\right) =& \mathrm{Var}\left(\frac{\sum_i^n (x_i - \overline{x}) Y_i}{\left(\sum_i^n x_i^2\right) - n\overline{x}^2}\right) \\
=& \sum_i^n \left(\frac{(x_i - \overline{x})}{\left(\sum_i^n x_i^2\right) - n\overline{x}^2}\right)^2 \mathrm{Var}\,(Y_i) \\
=& \frac{\sigma^2}{n\sum_i^n (x_i - \overline{x})^2}
\end{aligned}$$

More theorems about Linear Model estimators

- **Thm.:** Let $(x_1, Y_1), \ldots, (x_n, Y_n)$ satisfy the assumptions of the Linear Model. Then
  - $\hat{\beta}_1$, $\overline{\mathbf{Y}}$, and $\hat{\boldsymbol{\sigma}}^2$ are mutually independent.
  - $\frac{n}{\sigma^2}\hat{\boldsymbol{\sigma}}^2$ has a χ^2 distribution with $n-2$ degrees of freedom.
- **Corr.:** Let $\hat{\boldsymbol{\sigma}}^2$ be MLE for σ^2 in the Linear Model. Then
  - $\left(\frac{n}{n-2}\right)\hat{\boldsymbol{\sigma}}^2$ is an unbiased estimator for σ^2.
  - The random variables $\hat{\mathbf{Y}}^2$ and $\hat{\boldsymbol{\sigma}}^2$ are independent.
- Proofs relegated to appendix of Chapter 11 in Larsen & Marx text.

More theorems about Linear Model estimators

- Calculating $\sum_i^n (y_i - \hat{y}_i)^2 = \sum_i^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$

$$= \sum_i^n (y_i - \overline{y})^2 - \hat{\beta}_1^2 \sum_i^n (x_i - \overline{x})^2$$

$$= \sum_i^n y_i^2 - \frac{1}{n} \sum_i^n y_i - \frac{\left[\sum_i^n x_i y_i - \frac{1}{n}\left(\sum_i^n x_i\right)\left(\sum_i^n y_i\right)\right]^2}{\left(\sum_i^n x_i^2\right) - \frac{1}{n}\left(\sum_i^n x_i\right)}$$

$$= \sum_i^n y_i^2 - \hat{\beta}_0 \sum_i^n y_i - \hat{\beta}_1 \sum_i^n x_i y_i.$$

- Note that the above are true only for the actual estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, calculated from the data.
- The above are not true if $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_1$ are arbitrary.

Drawing inferences about $\hat{\beta}_1$

- **Thm.:** Let $(x_1, Y_1), \ldots, (x_n, Y_n)$ satisfy the assumptions of the Linear Model, and let

$$S^2 = \frac{1}{n-2} \sum_i^n \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 = \frac{1}{n-2} \sum_i^n \left(Y_i - \hat{Y}_i \right)^2$$

- Then the following has a Student T distribution with $n-2$ df.

$$T_{n-2} = \frac{\hat{\beta}_1 - \beta_1}{S / \sqrt{\sum_i^n (x_i - \overline{x})^2}}$$

Drawing inferences about $\hat{\beta}_1$

- **Pf.:** Note that the following is normally distributed

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_i^n (x_i - \overline{x})(Y_i - \hat{y}_i)}{\sum_i^n (x_i - \overline{x})^2}$$

- Note that, since $E(Y_i) = \hat{y}_i$, it follows $E\left(\hat{\beta}_1 - \beta_1\right) = 0$.
- Then note that

$$\text{Var}\left(\hat{\beta}_1 - \beta_1\right) = \frac{\sum_i^n (x_i - \overline{x})^2 \text{Var}(Y_i)}{\left[\sum_i^n (x_i - \overline{x})^2\right]^2} = \frac{\sigma^2}{\sum_i^n (x_i - \overline{x})^2}$$

- Hence the following is distributed as a standard normal

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{\sum_i^n (x_i - \overline{x})^2}}}$$

Drawing inferences about $\hat{\beta}_1$

- **Pf. (continued):** We know that

$$Z = \frac{\hat{\boldsymbol{\beta}}_1 - \beta_1}{\frac{\sigma}{\sqrt{\sum_i^n (x_i - \overline{x})^2}}}$$

  is distributed like a standard normal.

- We also know that

$$\frac{n}{\sigma^2}\hat{\boldsymbol{\sigma}}^2 = \frac{n-2}{\sigma^2}S^2$$

  is distributed as a χ^2 r.v. with $n-2$ df.

- Hence the following is T distributed with $n-2$ df

$$T_{n-2} = \frac{Z}{\sqrt{\frac{\left(\frac{n-2}{\sigma^2}S^2\right)}{n-2}}} = \frac{\hat{\boldsymbol{\beta}}_1 - \beta_1}{S/\sqrt{\sum_i^n (x_i - \overline{x})^2}} \quad \square$$

Drawing inferences about $\hat{\beta}_1$

- We are given data $(x_1, Y_1), \ldots, (x_n, Y_n)$ for Linear Model.
- Calculate the statistic

$$t = \frac{\hat{\beta}_1 - \beta_1'}{s/\sqrt{\sum_i^n (x_i - \overline{x})^2}}$$

where

$$s^2 = \frac{1}{n-2}\sum_i^n \left(y_i - \hat{\boldsymbol{\beta}}_0 - \hat{\boldsymbol{\beta}}_1 x_i\right)^2 = \frac{1}{n-2}\sum_i^n (y_i - \hat{y}_i)^2$$

- Hypothesis tests involving $\hat{\beta}_1$:
  - To test $H_0: \ \beta_1 = \beta_1'$ versus $H_1: \ \beta_1 > \beta_1'$ at α level of significance, reject H_0 if $t \geq +t_{\alpha,n-2}$.
  - To test $H_0: \ \beta_1 = \beta_1'$ versus $H_1: \ \beta_1 < \beta_1'$ at α level of significance, reject H_0 if $t \leq -t_{\alpha,n-2}$.
  - To test $H_0: \ \beta_1 = \beta_1'$ versus $H_1: \ \beta_1 \neq \beta_1'$ at α level of significance, reject H_0 if $t \geq +t_{\alpha/2,n-2}$ or $t \leq -t_{\alpha/2,n-2}$.

Confidence intervals for $\hat{\beta}_1$

- We are given data $(x_1, Y_1), \ldots, (x_n, Y_n)$ for Linear Model.
- As with hypothesis testing, calculate the statistic

$$t = \frac{\hat{\beta}_1 - \beta_1'}{s/\sqrt{\sum_i^n (x_i - \overline{x})^2}}$$

- Then

$$\left[\hat{\beta}_1 - \frac{t_{\alpha/2,n-2}\; s}{\sqrt{\sum_i^n (x_i - \overline{x})^2}},\;\; \hat{\beta}_1 + \frac{t_{\alpha/2,n-2}\; s}{\sqrt{\sum_i^n (x_i - \overline{x})^2}}\right]$$

  is a $100(1-\alpha)\%$ confidence interval for β_1.
- It is also possible to find confidence intervals for β_0, but usually not as important.

Covariance and correlation

- Recall the *covariance* of random variables X and Y

$$\mathrm{Cov}(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

- The covariance depends on the units of the variables.
- Make independent of the units by dividing by σ_X and σ_Y, to obtain the *correlation coefficient*,

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$$

- This also has the effect of ensuring $\rho(X,Y) \in [-1,+1]$.

Proof that $\rho(X, Y) \in [-1, +1]$

- Define the standardized r.v.s, $X^* = \frac{X-\mu_X}{\sigma_X}$ and $Y^* = \frac{Y-\mu_Y}{\sigma_Y}$
- Hence $E(X^*) = E(Y^*) = 0$ and $\mathrm{Var}(X^*) = \mathrm{Var}(Y^*) = 1$.
- Now consider

$$\begin{aligned} 0 \leq \mathrm{Var}(X^* \pm Y^*) =& E\left[(X^*)^2\right] + 2E(X^* Y^*) + E\left[(Y^*)^2\right] \\ =& \mathrm{Var}(X^*) \pm 2\mathrm{Cov}(X^*, Y^*) + \mathrm{Var}(Y^*) \\ =& 2 \pm 2\rho(X, Y). \end{aligned}$$

- It follows that $-1 \leq \rho(X, Y) \leq +1$. $\square$
- We proved above earlier using Cauchy-Schwarz inequality.

The Pearson correlation coefficient

- We have defined the correlation

$$\rho(X, Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

- So define the *sample correlation coefficient*,

$$R = \frac{\frac{1}{n}\sum_i^n X_i Y_i - \overline{X}\,\overline{Y}}{\sqrt{\frac{1}{n}\sum_i^n \left(X_i - \overline{X}\right)^2}\sqrt{\frac{1}{n}\sum_i^n \left(Y_i - \overline{Y}\right)^2}}$$

A simple relationship for r^2

- Now suppose that we have data $(x_1, y_1), \ldots, (x_n, y_n)$.
- Define the *coefficient of determination*

$$r^2 = \frac{\sum_i^n (y_i - \overline{y})^2 - \sum_i^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{\sum_i^n (y_i - \overline{y})^2}.$$

- Simple interpretation
  - $\sum_i^n (y_i - \overline{y})^2$ is the *total variability* in y.
  - $\sum_i^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$ is the variability that can not be explained by linear regression.
  - The numerator of r^2 is the variability that *can* be explained by linear regression.
  - The quantity r^2 is the fraction of the variability that can be explained by regression.

Summary

- We continued our study of the Linear Model.
- We showed that $\hat{\beta}_1$ is Student T distributed with $n-2$ degrees of freedom.
  - We used this to do hypothesis testing for $\hat{\beta}_1$.
  - We used this to construct confidence intervals for $\hat{\beta}_1$.
- We defined the *correlation* $\rho(X, Y) \in [-1, +1]$.
- We presented a method of estimating $\rho(X, Y)$ using sample moments.
- We constructed the *Pearson correlation coefficient* R.