Contents

1 Reading 9: Correlation and Regressions

1.1 Sample covar and sample correlation coefficients

Sample covariance: $cov_{x,y} = \sum_i \frac{(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$
Sample correlation coeff: $r_{x,y} = \frac{cov_{x,y}}{s_x s_y}$, where s_x is the sample dev of X.

1.2 Limtations to correlations analysis

Outliers: The results will be affected by extreme data points.(outliers)
Spurious correlation: There might be some non-zero corrlation coeff, but acutally they have no correlation at all.
Nonlinear relationships: Correlation only describe the linear relastions.

1.3 Hypothesis: determine if the population correlation coefficient is zero

Two-tailed hypothesis test:

$$H_0 : \rho = 0, H_a : \rho \neq 0$$

Assume that the two populations are **normally** distrubited, then we can use t-test:

$$t = \frac{r\sqrt{n-2}}{1-r^2}$$

: Reject H_0 if $t > +t_{critical}$ or $t < -t_{critical}$. Here,r is the sample correlation. Remember, you need to check t-table to find the t-value.

1.4 Determine dependent/indepedent variables in a linear regression

Simple linear regression: Explain the variation in a dependent variable in terms of the variabltion in a single indepedent variable. **Independent variables** are called explanatory variable, the exogenous variable, or the predicting variable. **Dependent variable** is also called the explained variable, the endogenous variable, or the predicted variable.

1.5 Assumptions in linear regression and interpret regression coeff.

1. Assumptions of linear regression:
   (a) Linear relationship must exist.
   (b) The indepedent variable is uncorrelated with residuals.
   (c) Expected Residual term is value. $E(\epsilon) = 0$
   (d) variance of the residual term is const. $E(\epsilon_i^2) = \sigma_\epsilon^2$
   (e) The residual term is independently distributed. $E(\epsilon_i \epsilon_j) = 1$
   (f) The residual term is normally distributed.
2. Simple Linear Regression Model
   (a) Model: $Y_i = b_0 + b_1 X_i + \epsilon_i$, where $i = 1...n$, and Y_i is the actual observed data.
   (b) The fitted line, the line of best fit : $\hat{Y} = \hat{b_0} + \hat{b}1 X_i$. Where $\hat{b_0}$ is the estimated parameter of the model.
   (c) How to choose the best fitted line? **Sum of squared errors** is minimum.

$$\hat{b_1} = \frac{cov_{x,y}}{sigma_x^2}$$

   where X is the indepdent variable. $\hat{b_1}$ is "regression coeffcient".

$$\hat{b_0} = \bar{Y} - \hat{b_1}\bar{X}$$

   where $\bar{X}, \bar{Y}$ are the mean.
3. Interpreting a regression coefficient: Similar to basic ideas of "slope". Keep in mind: any conclusion regarding this parameter needs the statistical significance of the slope coefficient.

1.6 Standard error of estimate, the coeff. of determination and a confidence interval for a regression coefficient.

1. Standard error of estimate (SEE): Standard deviation between $Y_{estimate}$ and Y_{actual}. - Smaller: better
2. Coefficient of Determination (R^2) The percentage of the total variance in the dependent variable that is predictable from the indepedent variable. - One indepdent variable: $R^2 = r^2$, where r^2 is the square of correlation coefficient.
3. Regression Coefficient confidence interval
   (a) Hypothesis: $H_0 : b_1 = 0 \Leftrightarrow H_a : b_1 \neq 0$
   (b) Confidence interval: $\hat{b_1} - (t_c s_{\hat{b_1}}) < b_1 < \hat{b_1} + (t_c s_{\hat{b_1}})$ $s_{\hat{b}_1}$ is the standard error of the regression coeffi.

1.7 Hypothesis: Determine if $\hat{b}_1 = b_1$

1. t-test statistic: $t_{b_1} = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}}$
2. Reject: if $t > +t_{critical}$ or $t < -t_{critical}$

1.8 Calculate the predicted value for the depedent variable

If an estimated regression model is known, $\hat{Y} = \hat{b}_0 + \hat{b}_1 X_p$

1.9 Calculate and interpret a confidence interval for the predicted value of the depedent variable

1. Eq: $\hat{Y} \pm (t_c s_f)$, where s_f is the std error of the forecast.

2. $s_f^2 = SEE^2 \left[1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{(n-1)s_x^2}\right]$

   (a) SEE^2 = variance of the residuals

   (b) s_x^2 = variance of the indepdent variable

   (c) X = value of the independent variable where the forecast was made.

1.10 ANOVA in regression. Interpret results, and calculate F-statistic

1. Analysis of variance (ANOVA) is used to analyze the total variability of the depedent variable.

   (a) Total sum of squares(SST): $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
   SST is the total variation in the depedent variable. $Variance = SST/(n-1)$

   (b) Regression sum of squares(RSS): $RSS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
   RSS is the explained variation.

   (c) Sum of squared errors(SSE): $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
   SSE is the unexplained variation.

   (d) $SST = RSS + SSE$ I cannot get this equation yet You need to know how to use these squares.

   (e) Degree of freedom: i) Regression(Explained): $k = 1$, since we only estimate one parameters. ii) Error(Unexplained) $df = n - k - 1 = n - 2$ iii) Total variation $df = n - 1$

2. Calculating R^2 and **SEE**

   (a) $R^2 = explainedvariation/totalvarn = RSS/SST$

   (b) $\mathbf{SEE} = \sqrt{\frac{\mathbf{SSE}}{\mathbf{n-2}}}$ **SEE** is the std deviation of the regression error terms.

3. The F-Statistic: used to explain whether *at least one* indepdent parameter can significanly explain the dependent parameter.

   (a) F-statistic eq: $F = \frac{MSR}{MSE} = \frac{RSS/k}{SSE/n-k-1}$ where MSR = mean regression sum of squares. MSE = mean squared errors. Note: One tailed test!

4. F-statistic with one independent variable.

   (a) Hypothesis: $H_0 : b1 = 0 \Leftrightarrow H_a : b1 \neq 0$

   (b) degree of freedom: $df_{rss} = k = 1, df_{sse} = n - k - 1$

   (c) Decision rule: reject H_0 if $F > F_c$

1.11 Limitations of regression analysis

1. Parameter instability: the estimation eq may not be useful for other times.

2. Limited usefulness: other participants may also use the same eq.

3. Assumptions does not hold: i) Heteroskedastic, i.e., non-const variance of the error terms. ii) autocorrelation, i.e., error terms are not independent.