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. $E(\epsilon_i^2) = \sigma_\epsilon^2$. Otherwise, it will be "heteroskedastic"
   (e) The residual term is independently distributed. otherwise – "auto correlation"" $E(\epsilon_i \epsilon_j) = 1$
   (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 : $\hat{Y} = \hat{b_0} + \hat{b}1X_i$. 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.

2 Reading 10: Multiple Regression and Issues in Regression Analysis

Some basic ides

1. Model: $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + ... + b_kX_{ki} + \epsilon_i$
2. Multiple regression methodology estimates the intercept and slope coefficients so that $\sum_i^n \epsilon_i^2$ is minimized.

2.1 Interpret estimated regression coefficients and their p-values.

They are just simple linear functions with multiple parameters. Ignore.

2.2 Formulate a null/alternative hypothesis, do correspoding calculations

1. Hypothesis Testing of Regression coefficient. (Multi-parameters).
   Use t-statistics to determine if one parameter significantly contribute to the model.

$$t = \frac{\hat{b}_j - b_j}{s_{\hat{b}_j}}, df = n - k - 1$$

   where k is the number of regression coefficients, and 1 corresponds to the intercept term, and $s_{\hat{b}_j}$ is the coefficient standard error of b_j

2. Determining statistical significance.
   "testing statistical significance" $\Rightarrow H_0 : b_j = 0, H_a : b_j \neq 0$

3. Interpreting p-values.

   (a) Def: p-value is **the smallest level of significance for which the null hypothesis can be rejected.** If the p-value is less than significance level, the null

4. Other Tests of the Regression Coefficients: $H_0 : a =$some value

2.3 Calculate and Interpret a confidence interval for the population value of a regression coefficient or a predicted value for the depedent variable if an estimated regression model.

1. Confidence intervals for a regress. coeff.: $\hat{b}_j \pm (t_c \times s_{\hat{b}_j})$

2. predicting the depedent variable: $\hat{Y}_i = \hat{b}_0 + \hat{b}_1\hat{X}_{1i} + ... + \hat{b}_k\hat{X}_{ki}$
   Even if you may conclude that some b_i are not statistally significantly, you cannot treat them as 0 and keep other parameters unchanged. You should use the original model, or you can throw $\hat{b}_k$ away and make a new regression model.

2.4 Assumptions of a multiple regresssion model

1. Linear relationships exist.

2. The independent variables are not random, and there is no exact linear relation between indepdent variables.

3. $E[\epsilon|X_1, ..., X_k] = 0$

4. Variance of $\epsilon = 0$, i.e. $E[\epsilon_i] = 0$

5. $E(\epsilon_i\epsilon_j) = 0$

6. ϵ is normally distributed.

2.5 Calculate and interpret F-statistic

F-test: whether at least **one** of the indepedent variables explains a significant portion of the variation of the depedent variable. F test is a one-tail test.

1. $H_0 : b1 = b2 = b3 = 0 vs H_a :$ at least one $b_j \neq 0$

2. $F = \frac{MSR}{MSE} = \frac{RSS/k}{SSE/n-k-1}$

3. Degree of freedom: $df_{numerator} = k, df_{denominator} = n - k - 1$

4. Rules: reject H_0 if $F(test - statistic) > F_c(criticalvalue)$

2.6 Distinguish between R^2 and adjusted R^2

1. coefficient of determination R^2: used to test if a group of indepedent variable can explain the depedent variable:
   $R^2 = \frac{totalvariation - unexplainedvariation}{totalvariation} = \frac{SST - SSE}{SST} = \frac{RSS}{SST}$
   Multiple R $= \sqrt{R^2}$

2. Adjusted R^2

   (a) Note: R^2: Overestimating: will increase as variables are added to the model. Even the marginal contribution of new variables are not statistically significant.

   (b) Introduce R_a^2: $R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right)\right](1 - R^2)$

2.7 Evaluate the quality of a regression model by analyzing the ouput of the equation/ANOVA table

1. ANOVA Tables, some important quantities

   (a) $R^2 = \frac{RSS}{SST}$

   (b) $F = \frac{MSR}{MSE}$ with k and $n - k - 1$ df

   (c) Standard error of estimate:$SEE = \sqrt{MSE}$

2.8 Formulate a multiple regression with dummpy variables to represent qualitative factors

1. Def: Some value is quite qualitative. Using dummy values like 0 or 1 to describe their impacts.

2. Note: Pay attention to # of dummy variables. If n classes, we must use $n - 1$ dummy variables.

3. Interpreting the coefficients in a dummy variable regression. We can use F-statistics to test a group of parameters, or use t-test to test the individual slope coefficients.

4. Example of Regression application with dummy variables. See Notes directly.

2.9 Why multiple regression isn't as easy as it looks?

Pay attention to the assumptions that have been used. Violations like::

1. Heteroskedasticity

2. Serial correlation (auto-correlation)

3. Multicollinarity

Any violations on the assumptions will impact the estimation of SEE, and finaly change the t-statistic and F-statistic, and change the conclustion of the hypotesis test.

2.10 Types of Heteroskedasticity, how heteroskedasiticity and serial correlation affect inference

1. What is Heteroskedasticity?
   Corresponding assumptions: Variance of the residuals is constant across observations. – **Homoskedasticity** Heteroskedasity means the variance of the residuals is not equal.

(a) Unconditional heter: Not related to the level of the indepedent variables. Will not systematiclly increase with changes in the value of the indepedent variables. Usually will not casue major problems.

(b) Conditional heter: Related to the level of the independent variables. Eg: Conditional heter exists if the variance of the residuals increase with the value of the independent variables increases. Will cause big problems.

2. Effect of Heteroskedasiticity on Regression Analysis

   (a) Unreliable standard errors.

   (b) The coefficient estimates aren't affected.

   (c) Will change the t-statistic, and will change the conclusion.

   (d) Unreliable F-test

3. Detect Heteroskedasicity

   (a) Scatter plot

   (b) Breusch-pagan test: $BPtest = n \times R^2_{resid}$ with $df = k$. where n =the number of observations, R^2_{resid}=R^2 from a second regression of the squared residuals from the first regression. k =the number of independent variables. If R^2 or BP-test are too large, something is wrong.

4. Correcting Heteroskedasticity

   (a) Calculate robust sndard errors(White corrected std errors.). Use them for t-test.

   (b) Generalized least squares.

5. What is serial correlations?

   (a) Def: auto-correlation, in which the residual terms are correlated. Common problem with time series data.

      i. Positive serial correlation: a postive error in one time period will increase the posibility to observe a positive one next time.

      ii. Negative serial correlation: Just opposite.

   (b) Effect: positive serial correlation will get small coefficient std errors. Thus, too large t-statistics. therefore, too many Type I errors: reject the null hypothesis H_0 while it's actually true.

   (c) Detection:

      i. Residual plots

      ii. Durbin-Watson statistics:

$$DW = \frac{\sum_{t=2}^{T}(\hat{\varepsilon}_t - \hat{\epsilon}_{t-1})^2}{\sum_{t=1}^{T}\hat{\epsilon}^2}$$

For large samples, $DW \approx 2(1 - r)$, where r is the correlation coefficient between residuals from one period and thoese from the previous period.

Results:

A. $DW = 2 \Rightarrow$ Homoskedasitic and not serially correlated.

B. $DW < 2 \Rightarrow$ Postively serially correlated.

C. $DW > 2 \Rightarrow$ Negatively serially correlated.

Formulated hypothesis with DW-table, upper and lower critical values

A. Hypothesis: H_0 : the regression has **no** positive serial correlation.

B. $DW < d_l$: positive serially correlated. Reject null.

C. $d_l < DW < d_u$: inconclusive results.

D. $DW > d_u$: **There is no evidence that are positive correlated.**

(d) Correcting serial correlation:

i. Adjust the coefficient std errors. **recommended.** Using Hansen method.

A. Serial correlation only: Hansen method.

B. Heteroskedasticity only: White-corrected stand errors.

C. Both: Hans methods.

ii. Imporoe the specification of the model.

2.11 Multicolinearity and its cause and effects in regression analysis

Multicollinearity: Indepedent variables or linear combinations of independent variables are highly correlated.

1. Effect of Multicollinearity on Regression Analysis: Will increase the std errors of the slope coefficients. Type II Error: A variable is significant, while we conclude it's not.

2. Detecting: Common situation: $t - statistic$ is not significant while $F - test$ is significant. This tells us the indepedent variables are highly correlated.
A simple rule works if there are 2 indepedent variables: when the absolute value of the sample correlation betewen any two indepedent variables in the regression is greater than 0.7.

3. Correcting: omit one or more of the correlated indepedent variables. THe problame is that it's hard to find the variables that result in the multicolinearity.

2.12 Model misspecification

1. Defination of **Regression model of specification**: decide which independent variables to be included in the model.

2. Types of misspecification

(a) The functional form can be misspecified: important variables are ommited;variables should be transformed; data is improperly pooled.

(b) Explanatory variables are correlated with error term in time series model: A lagged dependent variable is used as an independent variable; a function of the dependent variable is used as an independent variable(forecasting the past); independent variables are measured with error.

(c) Other time-series misspecification.

2.13 Models with qualitive dependent variables

Include qualitative dependent variables, like default, bankcrupcy. Cannot use an ordinary regression model. Should use other models like **probit and logit models** or **discriminant models**.

1. Probit: normal distribution, give probability.

2. Logistic: logistic distribution.

3. Discriminant: result in an overall score or ranking.

3 Reading 11: Time-Series Analysis

3.1 Calculate/evaluate the predicted trend value for a time series given the estimated trend coefficients

1. Linear Trend Model and Log-linear Trend

(a) Definition: $y_t = b_0 + b_1(t) + \epsilon_t$ Note: t is just time.

(b) Coefficients is determined by OLS. Ordinary least squared regression.
$\hat{y} = \hat{b}_0 + \hat{b}_1$

(c) Log-linear Trend Models

(d) Model: $y_t = \exp b_0 + b_1(t) \Rightarrow \ln y_t = b_0 + b_1(t)$

3.2 Factors that determine whether a linear or a log-linear model trend should be used

1. Factors that determine which model is best: plot data.

2. Limitaions of trend models:

   (a) residuals are uncorrelated with each other. Otherwise, it will cause auto correlation and we should not use the trend model.

   (b) For log-linear model, it is not suitable for cases with serial correlations (autocorrelation).

   (c) Detect auto correlation: Durbin Watson statistic. $DW = 2.0 \Rightarrow$ No auto correlation.

3.3 Autoregressive model, requirements for covariance statinoary

1. Autoregressive model:

   (a) Model: $x_t = b_0 + b_1 x_{t-1} + \varepsilon_t$

   (b) Statistical inferences bse on ordinary least squares estimates doesn't apply unless the time series is **covariance stationary**.

   (c) Conditions for covariance stationary

      i. Constant and finite expected value.

      ii. Constant and finite variance.

      iii. Constant and finite covariance between values at any given lag.

3.4 An autogressive model of order p

1. Model(order p): $x_t = b_0 + b_1 x_{t-1} + b_2 x_{t_2} + ... + b_p x_{t-p} + \varepsilon_t$

2. Forecasting with an autoregressive model:

   (a) One-period-ahead forecast for $AR(1)$: $\hat{x}_{t+1} = \hat{b}_0 + \hat{b}_1 x_t$

   (b) Two-period-ahead forecast for $AR(1)$: $\hat{x}_{t+2} = \hat{b}_0 + \hat{b}_1 \hat{x}_{t+1}$

3.5 How the residuals can be used to test the autogressive model

1. The residual should have no *serial correlation* if an AR model is correct.

2. Steps

   (a) Estimate: Start with AR(1)

   (b) Calculate: the autocorrelations of he model residuals

   (c) Test: whether the autoccorelations are signficantly different from 0.
   The standard error is $\frac{1}{\sqrt{T}}$ for T observations. The t-test for each observation is $t = \frac{\rho_{\epsilon_t, epsilon_{t-k}}}{1/sqrtT}$, with $T-2$ df.

3.6 Mean reversion and a mean-reverting level

1. Mean reversion: The time series tends to move toward its mean.
2. Mean-reverting level: $\hat{x}_{t+1} = x_t$, where $\hat{x}_t$ is the predicted value.
3. All covariance stationary time series has finite mean-reverting level.

3.7 Contrast in-sample and out-of-sample forecasts and the forecasting accuracy of different time-series models based on the root mean squared error criterion.

1. in-sample, out-of-sample: determined by if the predicted data is in the range of the observations.
2. RMSE, root mean squared error: used to compare the accurancy. If the accurancy of out-of-sample is better, you should use it for future applications

3.8 Explain the instability of coefficients of time-series models

1. Instability or nonstationarity. Due to the dynamic econimic conditions, model coefficients will change a lot from period to period.
2. Shorter time series are more stable, but longer time series are more reliable.

3.9 Random walk processes and their comparisons between covariance stationary processes

1. Random walk: $x_t = x_{t-1} + \varepsilon_t$
   (a) $E(\varepsilon_t) = 0$: The expected value of each error is zero.
   (b) $E(\varepsilon_t^2) = 0$: The variance of the error terms is constant.
   (c) $E(\varepsilon_i, \varepsilon_j) = 0$: There is no serial correlation in the error terms.
2. Random walk with a Drift: $x_t = b_0 + b_1 x_{t-1} + \varepsilon_t$, where b_1=0
3. A random walk or a random walk with a drift have no finte mean-reverting level. Since $b_1 = 1$, $\frac{b_0}{1-b_1} = \frac{b_0}{0}$. Therefore, they are not covariance stationary.
4. $b_1 = 1$, they exhibit a unit root. Thus, **the least square regression that been used in AR(1) will not work unless we transfrom the data**.

3.10 Things about unit roots: when they will occur, how to test them, how to transform data to apply AR

1. Unit root testing for nonstationarity:
   (a) run an AR model and check autocorrelations
   (b) perform Dickey Fuller test.
      i. Transform: $x_t = b_0 + b_1 x_1 + \varepsilon \Rightarrow x_t - x_{t-1} = b_0 + (b_1 - 1)x_{t-1} + \varepsilon$
      ii. Direct test if $b_1 - 1 = 0$ using a modified t-test.
2. First differencing
   (a) For a random walk, transform the data $y_t = x_t - x_{t-1} \Rightarrow y_t = \varepsilon_t$ then start to use an AR model $y = b_0 + b_1 y_{t-1} + \varepsilon$, where $b_0 = b_1 = 0$
   (b) y is covariance stationary.

3.11 How to test and correct for seasonality in a time-series model, and calculate and interpret a forecasted value using an AR model with a sesonal lag.

1. Detect: special autocorrelation exists for some seasonal lags.
2. Correction: Add an additional seasonal lag term.

3.12 Explain autogressive conditional heteroskedasticity (ARCH) and describe how ARCH models can be applied to predict the variance of a time series

1. ARCH: the variance of the residuals in one period is dependent on the variance of the residuals in a previous period.
2. Using ARCH models:
   Example $ARCH(1)$: $\hat{\varepsilon}_t^2 = a_0 + a_1\hat{\varepsilon}_{t-1} + \mu_t$ if a_1 is significantly different from zero. $\hat{\varepsilon}_t^2$ is the squared residuals.
   Note: Things like generalized least squares should me used to correct heteroskedasticity. otherwise, the std errors of the coefficients will be wrong, leading to invalid conclusions.
3. Predicting the variance of a time series: using ARCH model to predict the variance of future periods: $\hat{\sigma}_{t+1}^2 = \hat{a}_0 + \hat{a}_1\hat{\varepsilon}_t^2$