Statistics 5350/7110
Forecasting

Lecture 5
Multiple Linear Regression

Professor Stine

Admin Issues

- Questions?

- Assignments
  - Questions about A1 (due Thursday)

- Review
  - Estimated autocovariance and autocorrelation

$$\widehat{\gamma}_x(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_{t+h} - \overline{X})(X_t - \overline{X}) \qquad \widehat{\rho}_x(h) = \frac{\widehat{\gamma}_x(h)}{\widehat{\gamma}_x(0)}$$

  - Properties of estimates and effects of dependence
  - Like all correlations, autocorrelations influenced by outliers and misses nonlinearity (Lecture_4.Rmd)
  - Nonstationarity and spurious correlation
  - Cross-correlation

Today's Topics

Textbook §3.1, 3.2

- Multiple regression model
  - Step back from time series for this and the next class
  - Assumptions
  - Least squares estimates
  - Inference
- Collinearity
  - Effect of correlated explanatory variables, common with time series
  - Variance inflation factor (VIF) and impact of collinearity on inference
- Prediction
  - Prediction interval versus confidence interval
  - Extrapolation effect
  - Role of normality
- Model selection criteria
  - AIC, BIC, and RIC

Concepts should be familiar from other courses on regression analysis

Multiple Regression Model

- Model combines an equation with assumptions for deviations
  - Model for a time series X_t with q predictors $Z_1, Z_2, \ldots, Z_q$
  - Equation is weighted sum ("linear combination")

$$X_t = \beta_0 + \beta_1 Z_{t,1} + \cdots + \beta_q Z_{t,q} + w_t$$

  - White noise errors: mean zero, common variance σ_w^2, ideally normally distributed
- Least squares estimates
  - Minimum variance, unbiased estimates

$$\widehat{\beta} = \arg\min_{\beta} \sum_{t=1}^{n} (X_t - \beta_0 - \beta_1 Z_{t,1} - \cdots \beta_{t,q})^2$$

  - Formula in simple regression ($q = 1, Z_{t,1} = Z_t$) suggests general form (and so should be known)

$$\hat{\beta}_1 = \frac{\sum_t (Z_t - \overline{Z})\, X_t}{\sum_t (Z_t - \overline{Z})^2} = \frac{\sum_t (Z_t - \overline{Z})\, X_t}{SS_Z} = \sum_{t=1}^{n} c_t\, X_t$$

where the weights are $c_t = (Z_t - \overline{Z})/SS_Z$.

Choice of symbols matches those in the textbook

Inference

- Fitted values and residuals
  - Fitted values, $\hat{X}_t = \widehat{\beta}_0 + \widehat{\beta}_1 Z_{t,1} + \cdots + \widehat{\beta}_q Z_{t,q}$
  - Residuals are deviations from fitted values, $\hat{w}_t = X_t - \hat{X}_t$
  - Sums of squares: $SST = \sum (X_t - \overline{X})^2,\ SSE = \sum (X_t - \hat{X}_t)^2 = \sum \hat{w}_t^2,\ SSR = SST - SSE$
  - Unbiased estimator of noise variance, $s_w^2 = SSE/(n - q - 1)$

d.f.

Names of sums of squares are idiosyncratic from book-to-book

- Standard errors
  - Estimated SD of the sampling distribution of an estimator.
  - Example when q=1

$$\mathrm{se}(\widehat{\beta}_1) = \frac{s_w}{\sqrt{SS_Z}} = \frac{\text{SD of residuals}}{\sqrt{n-1}\,(\text{SD of predictor})} \approx \frac{1}{\sqrt{n}}\,\frac{\text{variation around fit}}{\text{variation of predictor}}$$

- Tests
  - t-test and confidence interval for a single coefficient
  - F-test for more than one coefficient
  - Valid for moderate sample size when white noise w_t is not normal, but must be independent

Inference Details

- Single coefficient
  - Test $H_0 : \beta_j = 0$ with $t = \dfrac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \sim t_{n-q-1}$ when null hypothesis holds
  - Confidence interval $[\hat{\beta}_j \ \pm \ t_{\alpha/2,n-q-1}\text{se}(\hat{\beta}_j)]$ where $t_{.025,n-q-1} \approx 2$ when $\alpha = 0.05$
- Test of all coefficients
  - Test using F ratio (signal to noise ratio)

$$F = \frac{SSR/q}{SSE/(n-q-1)} = \frac{MSR}{MSE}$$

  Reject $H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0$ if $F > F_{\alpha,q,n-q-1}$

- Test of subset of coefficients
  - Test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_r = 0$ for $r \leq q$ using

aka, partial F test

$$F = \frac{(SSR - SSR_r)/r}{MSE}$$

where SSR_r is the regression SS of the restricted model

Inference Details

- R-squared statistic
  - Proportion of total sum-of-squares captured by model

$$R^2 = \frac{SSR}{SST} \qquad\qquad 1 - R^2 = \frac{SSE}{SST}$$

  - Square of the usual correlation corr(X_t, Z_t) when q =1 , square of corr$(X_t, \hat{X}_t)$ in general
- All coefficients
  - F ratio (signal to noise ratio)

$$F = \frac{SSR/q}{SSE/(n-q-1)} = \frac{R^2/q}{(1-R^2)/(n-q-1)}$$

- Subset of r coefficients (partial F)
  - Test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_r = 0$ for $r \leq q$ using

$$F = \frac{(\text{change in } R^2)/r}{(1-R^2)/(n-q-1)}$$

Collinearity

- Interpretation of a regression coefficient
  - Simple regression: difference in averages, marginal correlation
  - Multiple regression: difference in averages when comparable on other predictors, partial correlation

partial correlation important concept in time series models

- Collinearity
  - Correlation among explanatory variables, a.k.a. predictors, $Z_1, Z_2, \ldots Z_q$ in the model

common in time series models with lags of a variable

- Signs of collinearity
  - Coefficients change signs as predictors enter/leave model, "difficult to interpret"
  - Estimates have large standard errors

Uncorrelated

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma_w^2}{SSZ_j}$$

Correlated

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma_w^2}{SSZ_j \text{ given other predictors}} = \frac{\sigma_w^2}{(1 - R_j^2)\ SSZ_j}$$

where R_j^2 is the R-squared statistic of regressing Z_j on $Z_{k \neq j}$

- The increase in sampling variation is called the variance inflation factor $\quad \text{VIF}_j = 1/(1 - R_j^2)$

Prediction

- Predicted value
  - Plug known z's into fitted equation $\hat{X} = \hat{\beta}_0 + \hat{\beta}_1 z_1 + \hat{\beta}_{s,2} z_2 + \cdots + \hat{\beta}_q z_q$
- Confidence interval for mean
  - Target for inference is average response, $E(X \mid Z_1, \ldots, Z_q) = \beta_0 + \beta_1 z_1 + \cdots + \beta_q z_q$
  - Special case when q = 1

$$\text{Var}\left((\beta_0 + \beta_1 z) - (\hat{\beta}_0 + \hat{\beta}_1 z)\right) = \sigma_w^2 \left(\frac{1}{n} + \frac{(z - \overline{Z})^2}{SS_Z}\right)$$

extrapolation penalty

Confidence interval is the estimated value plus/minus ≈ 2 times the square root of the estimated variance

- Prediction interval for response value
  - Target for inference is response for one case, $X = \beta_0 + \beta_1 z_1 + \cdots + \beta_q z_q + w$
  - Special case when q = 1

$$\text{Var}(X - \hat{X}) = \text{Var}\left((\beta_0 + \beta_1 z + w) - (\hat{\beta}_0 + \hat{\beta}_1 z)\right) = \sigma_w^2 \left(1 + \frac{1}{n} + \frac{(z - \overline{Z})^2}{SS_Z}\right)$$

Residuals vs Prediction Errors

- Distinction
  - Residual: deviation of X_r from fitted value $\hat{X}_r$ for data used to fit model (1 ≤ s ≤ n)
  - Prediction error: deviation from predicted value for data X_s **not** used to fit model
- So what?
  - Suppose the model has no intercept and $\overline{Z} = 0$. Then
$$\hat{\beta} = \sum (Z_t X_t)/SS_Z = \sum \big(Z_t(w_t + \beta Z_t)\big)/SS_Z = \beta + \sum w_t Z_t/SS_Z$$
  - Variance of a residual is <u>smaller</u> than σ_w^2
$$\text{Var}(X_r - \hat{X}_r) = \text{Var}(w_r + (\beta - \hat{\beta})Z_r) = \sigma_w^2 + Z_r^2\text{Var}(\hat{\beta}) - 2\,Z_r\,\text{Cov}(w_r, \hat{\beta})$$
$$= \sigma_w^2\left(1 + \frac{Z_r^2}{SS_Z} - 2\,\frac{Z_r^2}{SS_Z}\right) = \sigma_w^2\left(1 - \frac{Z_r^2}{SS_Z}\right)$$
  - Variance of prediction error is <u>larger</u> than σ_w^2
$$\text{Var}(X_s - \hat{X}_s) = \sigma_w^2\left(1 + \frac{Z_s^2}{SS_Z}\right)$$

Same formula, different target

Even if model errors have equal variance, residuals almost never have equal variance

Implications

- Variance of residual depends on position
  - Even if the model errors have equal variance
- Need to adjust residual sum-of-squares when estimating σ_w^2
  - In case of the simple no-intercept, one-predictor model (prior slide)

$$E(SSE) = E\left(\sum \widehat{w}_t^2\right) = \sigma_w^2\left(n - \sum \frac{Z_t^2}{SS_Z}\right) = (n-1)\,\sigma_w^2$$

  - If the model has q explanatory variables, then $E(SSE) = (n-q-1)\sigma_w^2$
  - Hence we use the unbiased estimator $s_w^2 = SSE/(n-q-1)$ for σ_w^2
- Prediction sum-of-squares grows with the number of predictors
  - Suppose we predict an independent copy of the response
  - Model has q explanatory variables

$$E\left(\sum \widetilde{w}_t^2\right) = (n+q+1)\sigma_w^2$$

Concept of leverage covered in next class

Model Selection Criteria

- Idea
  - Need better objective than "maximize R^2" since this approach would use every possible predictor.
- Adjust for estimation error
  - Pick the set of explanatory variables that maximizes an estimate of the prediction error
  - The expected value of the residual SS is too small on average $E(SSE) = (n - q - 1)\,\sigma_w^2$
  - Adding predictor increases the expected prediction error (defined previously)

$$E\left(\sum \widetilde{w}_t^2\right) = (n + q + 1)\,\sigma_w^2$$

  - Hence $(n + q + 1)\,s_w^2$ is an unbiased estimate of the squared prediction error.
- Akaike information criterion
  - Let k denote the total number of estimated parameters (e.g. k = q + 1)
  - Define the biased estimator $\hat{\sigma}_k^2 = SSE(k)/n$
  - Then $\dfrac{n(n+k)}{n-k}\hat{\sigma}_k^2$ is unbiased. AIC defined similarly, AIC$(k) = \log \hat{\sigma}_k^2 + \dfrac{n+2k}{n}$

MLE estimator

Definition 3.2
Note footnote on page 41

Further Discussion of AIC

- AIC details
  - Likelihood is "probability of data" Y given parameters denoted θ, P(Y | θ)
  - Estimates of the parameters θ in a model commonly chosen to maximize the likelihood (MLE)
  - In a least squares regression fit to normally distributed data,

$$\max_\theta \log P_\theta(Y_1, \ldots, Y_n) = -\frac{n}{2}\left(1 + \log(2\pi\,\hat\sigma_k^2\right)$$

Covered in greater depth in other courses

  - For AIC, adding parameters must improve likelihood by enough to overcome a penalty

$$\text{AIC}(k) = -2\max_\theta \log P_\theta(Y) + 2\,k = n\,\log\hat\sigma_k^2 + 2\,k + c_n$$

    where c_n is a constant that depends on n (and so doesn't influence the choice of a model)
  - Expressions in text hide c_n and compute an average by dividing by n.
- Penalized likelihood
  - Overcomes the problem of maximizing R^2 which always increases
  - How much penalty is the issue: AIC chosen to give unbiased estimate of likelihood
  - Reasonable if the motivating probability model is roughly correct:

    Normal with independent observations

Model Selection Criteria

- Textbook version
  - AIC (without the distracting 1) $\text{AIC}(k) = \log \hat{\sigma}_k^2 + \frac{2k}{n}$
  - Corrected AIC $\text{AIC}_c(k) = \log \hat{\sigma}_k^2 + \frac{n+k}{n-k-2}$
  - Bayesian information criterion (BIC) $\text{BIC}(k) = \log \hat{\sigma}_k^2 + \frac{k \log n}{n}$
  - Risk inflation criterion (RIC) $\text{RIC}(k) = \log \hat{\sigma}_k^2 + \frac{k \log Q}{n}$

  where Q denotes the number of possible features available to use in the regression

divide by n

Definition 3.2-3.4

- Comparison
  - Different ways to penalize for the number of parameters in the model:
    AIC is most liberal, BIC is more restrictive (larger penalty). RIC is often much more strict.
  - AIC originated in search for number of autoregressive lags
  - BIC is "consistent": If the "true model" lies within the search space, BIC will choose it (eventually)
  - RIC penalizes for the scope or the search and is essentially Bonferroni selection

Examples

- Regression inference in R
  - Illustrate t and F tests
- Polynomials and collinearity
  - Estimate non-stochastic trends
  - Use of centering to reduce collinearity
- Model selection calculations
  - Showing how R computes these statistics

What's next?

- More multiple regression!
  - Diagnostics, emphasizing plots
  - Overall model fit: calibration and residual plots

    Calibration plot

    Residuals and quantile plots

    Leverage and outliers
  - Added variable plots

- Further examples and discussion of regression
  - Two handouts on regression in Notes folder on Canvas
  - Conceptual, doesn't use R for calculations