Linear Regression

Hui Chen

MIT Sloan

15.457, Spring 2021

Outline

Motivation: Modeling Stock Returns

- Market model

$$R^e_{i,t} = \alpha + \beta R^e_{m,t} + \varepsilon_{i,t}$$

- Multi-factor model

$$R^e_{i,t} = \alpha + \beta_1 f_{1,t} + \cdots + \beta_K f_{K,t} + \varepsilon_{i,t}$$

- Predicting returns

$$R_{m,t+1} = a + b\frac{D_t}{P_t} + \varepsilon_{t+1}$$

- What do these models have in common?
- Why might we be interested in studying these models?

Example: Return Predictability

Cochrane (JF 2011)

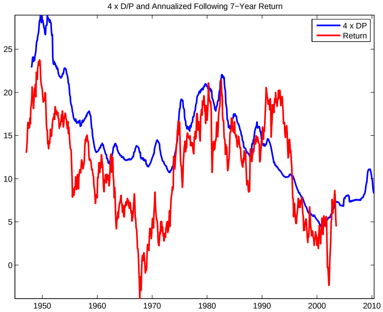


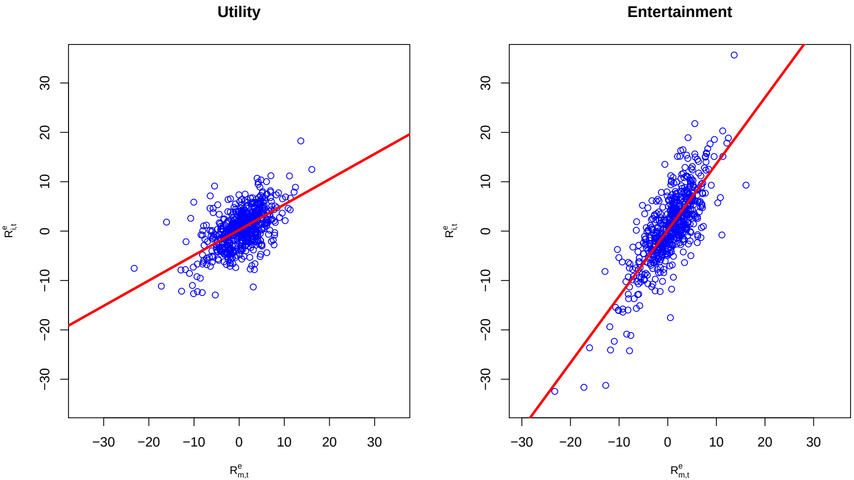
Table I

Return-Forecasting Regressions

The regression equation is $R^e_{t\to t+k} = a + b \times D_t/P_t + \varepsilon_{t+k}$. The dependent variable $R^e_{t\to t+k}$ is the CRSP value-weighted return less the 3-month Treasury bill return. Data are annual, 1947–2009. The 5-year regression t-statistic uses the Hansen–Hodrick (1980) correction. $\sigma[E_t(R^e)]$ represents the standard deviation of the fitted value, $\sigma(\hat{b} \times D_t/P_t)$.

Horizon k	b	$t(b)$	R^2	$\sigma[E_t(R^e)]$	$\frac{\sigma[E_t(R^e)]}{E(R^e)}$
1 year	3.8	(2.6)	0.09	5.46	0.76
5 years	20.6	(3.4)	0.28	29.3	0.62

Example: Market Model



Regression Statistics

R output (Entertainment Industry Portfolio)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.20271    0.20309   0.998    0.319
MktRF        1.34270    0.04469  30.048   <2e-16 ***
---

Residual standard error: 4.838 on 574 degrees of freedom
Multiple R-squared:  0.6113,Adjusted R-squared:  0.6107
F-statistic: 902.9 on 1 and 574 DF,  p-value: < 2.2e-16
```

- $\hat{\beta}_1 =$
- $SE(\hat{\beta}_1) =$
- t-statistic $= \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} =$
- p-value
- 95% conf. interval for β_1:

- $R^2 = 1 - \frac{RSS}{TSS} =$
- Residual standard error $RSE = \sqrt{\frac{1}{n-2} RSS} =$

Outline

Multiple Linear Regression

- Data: $(y_i, x_{i1}, \cdots, x_{ip}),\ i = 1, \cdots, n$
- Model:

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

  $\hookrightarrow$ What about the intercept?

- Matrix notation:

$$Y = X\beta + \varepsilon$$
$$y_i = \mathbf{x}_i'\beta + \varepsilon$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}, \quad \mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- How (not) to interpret the coefficients?

$$\beta_j = \frac{\partial E(y_i | x_{i1}, \cdots, x_{ip})}{\partial x_{ij}}$$

Multiple Linear Regression

Assumptions

1. Linearity: $Y = X\beta + \varepsilon$
2. Full rank: X is an $n \times p$ matrix with rank p. (identification condition)
3. Exogeneity of the independent variables: $E[\varepsilon_i|X] = 0$
4. Homoscedasticity and nonautocorrelation: $E[\varepsilon\varepsilon'|X] = \sigma^2\mathbf{I}$

Least Squares Estimator: Derivation

- Find β that minimizes the RSS:

$$\min_{\beta} \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta) = Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

- FOC:

$$-2X'Y + 2X'X\hat{\beta} = 0 \quad \Rightarrow \quad \hat{\beta} = (X'X)^{-1}X'Y$$

- Full rank condition for X ensures unique solution to least square problem (check second derivative).
- Asymptotic distribution (i.e., when n is large) of $\hat{\beta}$:

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon$$

$$\hat{\beta} - \beta \overset{a}{\sim} N\left(0, \sigma^2 (X'X)^{-1}\right) \qquad (CLT)$$

LS estimator for multiple regression

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$Var[\hat{\beta}|X] = \sigma^2 (X'X)^{-1}$$

Least Squares Estimator: Variance of the estimator

- The least squares estimator is **BLUE** (best linear unbiased estimator).
  - ↪ "Best" in the sense that it has the minimum variance among all linear *unbiased* estimators (Gauss-Markov Theorem).
  - ↪ Linear estimators: $\tilde{\beta} = CY$
  - ↪ A biased estimator could have even smaller variance (bias-variance tradeoff).

- Estimating σ^2:

$$\hat{\sigma}^2 = \frac{RSS}{n-p}$$

where

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2 = \hat{\varepsilon}'\hat{\varepsilon}$$

- Heteroscedasticity: $E[\varepsilon\varepsilon'] = \Omega$

$$\hat{\beta} - \beta \overset{a}{\sim} N\left(0, (X'X)^{-1} X'\Omega X (X'X)^{-1}\right)$$

- How to estimate $\widehat{\Omega}$? More on this later.

Regression Statistics

- Goodness of fit measures
  - Residual standard error (RSE)

$$RSE = \hat{\sigma} = \sqrt{\frac{RSS}{n-p}}$$

  - R^2 statistic

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

  - Adjusted R^2

$$\overline{R}^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}$$

- Significance of coefficients
  - t-statistic: $t = \dfrac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$, with $n-p$ degrees of freedom.
  - p-value: probability of observing a value equal to or above $|t|$, assuming $\beta_j = 0$
  - Confidence interval: $\left[\hat{\beta}_j - t_{\alpha/2} SE(\hat{\beta}_j), \hat{\beta}_j + t_{\alpha/2} SE(\hat{\beta}_j)\right]$
  - F-statistic: Does any of the (non-constant) predictor show significant effects?

$$F = \frac{(TSS - RSS)/(p-1)}{RSS/(n-p)}$$

Outline

Multicollinearity

- If the predictor variables are independent, the LS estimates from the multiple linear regression will be the same as obtained by separate simple regressions.
- In such cases, holding σ^2 fixed, more variability in the feature variables reduces the standard errors for $\hat{\beta}$.
- Multicollinearity: When two or more predictors are closely related, the accuracy of the least square estimates is substantially reduced.
- To diagnose multicollinearity, compute the variance inflation factor (VIF)

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

$R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors

$$X_j = X_{-j}\gamma + \varepsilon$$

Multicollinearity: VIF

- To see why multicollinearity reduces accuracy, consider an example with two de-meaned features ($\hat{E}[X_1] = \hat{E}[X_2] = 0$):

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

  ↪ LS estimator:

$$\text{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

$$= \frac{\sigma^2}{n} \frac{1}{\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\sigma}_{12}^2} \begin{bmatrix} \hat{\sigma}_2^2 & -\hat{\sigma}_{12} \\ -\hat{\sigma}_{12} & \hat{\sigma}_1^2 \end{bmatrix}$$

  ↪ Notice that $\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\sigma}_{12}^2} = \frac{1}{\hat{\sigma}_1^2} \frac{1}{1 - \frac{\hat{\sigma}_{12}^2}{\hat{\sigma}_1^2\hat{\sigma}_2^2}} = \frac{1}{\hat{\sigma}_1^2} VIF(\hat{\beta}_1)$

  ↪ Special case: Independent feature variables $\sigma_{12} = 0$

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2}{n} \begin{bmatrix} \frac{1}{\hat{\sigma}_1^2} & 0 \\ 0 & \frac{1}{\hat{\sigma}_2^2} \end{bmatrix}$$

Misspecification

- So far we have been assuming the correct specification of the linear model is known.
- Two most common specification errors in regression models:
  1. Omission of relevant variables.
  2. Inclusion of irrelevant variables.
- Omission of relevant variables typically causes the LS estimator to become *biased*, unless the omitted variables are uncorrelated or have no effects on y.
- When irrelevant variables are included, the LS estimator is still *unbiased*.
  - ↪ Intuition:
- This does not mean we should "overfit" the model by including many features!
  - ↪ Q: Why not?
- More on variable selection (forward, backward, mixed ...) later.

Misspecification: Omitted Variables

- Suppose the correctly specified model is

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

- Instead, we estimate the model with only X_1.

$$b_1 = (X_1'X_1)^{-1}X_1'Y =$$

- This leads to the **omitted variable formula**:

$$E[b_1|X] = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2$$

Bias exists unless $\beta_2 = 0$ or $X_1'X_2 = 0$.

- For example, we might overstate the effect of X_1 if ...

Misspecification: Example

CEO compensation

- As financial consultant, we want to examine the determinants of CEO compensation across firms in order to advise clients on the design of compensation packages.
- Suppose we use the following model:

$$y_i = \beta_0 + \beta_1 SIZE_i + \beta_2 EDU_i + \cdots + \varepsilon_i$$

  - ↪ y_i: measure of executive compensation
  - ↪ $SIZE_i$: firm size
  - ↪ EDU_i: measure of executive education level

- It is very difficult to measure the managerial ability of an executive. Education is at best a very noisy proxy.
- How would the omission of managerial ability affect the coefficient on firm size β_1?
- Q: Should you be concerned with such biases?

Other Considerations

- Influential outliers
  - ↪ Is it data error or informative observation?
- Heteroskedasticity
  - ↪ Plot the absolute residuals against the predicted responses ($|\hat{\varepsilon}_i|$ vs. $\widehat{y}_i$) and look for systematic trend.
  - ↪ Need to correct for the standard errors or use weighted least squares.
- Nonlinearity
  - ↪ Plot the residuals against the predictors and look for any nonlinear trend.
  - ↪ To fix the issue, consider adding nonlinear terms in the predictors, transform the response variables (e.g., Box-Cox transformation), or transform both sides. (More on this later.)
- Nonstationary
  - ↪ Is it a good idea to use stock price to predict monthly returns?

Outline

Qualitative Predictors

- Example: When predicting credit scores, *credit card balance* is a quantitative predictor; *student status* is a qualitative predictor.
- Use dummy variables to model qualitative predictors (e.g., $x_i = 1$ for student; 0 otherwise).

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \varepsilon_i & \text{if } i\text{th person is not a student} \\ \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is a student} \end{cases}$$

- Interpretation of β_0 and β_1.
- Qualitative predictors with $n > 2$ levels: Use $n-1$ dummies $x_{i1}, \cdots, x_{i,n-1}$.
  - $\hookrightarrow$ Q: Why not n?

Interactions

- We can capture certain nonlinear effects by adding interactions and nonlinear terms.
- Example:

$$R_{m,t+1} = a + b \ln \left(\frac{D}{P} \right)_t + \varepsilon_{t+1}$$

- We might suspect the predictive power of dividend yield to change depending on market volatility (use VIX as a proxy).

$$R_{m,t+1} = a + b \ln \left(\frac{D}{P} \right)_t + c\, VIX_t + d \ln \left(\frac{D}{P} \right)_t VIX_t + \varepsilon_{t+1}$$

- Interpretation:

$$R_{m,t+1} = a + \underbrace{(b + d\, VIX_t)}_{b(VIX_t)} \ln \left(\frac{D}{P} \right)_t + c\, VIX_t + \varepsilon_{t+1}$$

Hierarchical principle

If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

Nonlinearity

- More general nonlinear regression model:

$$y_i = f(\mathbf{x}_i; \beta) + \varepsilon_i$$

  - $\hookrightarrow$ $f()$ is a known function, with unknown parameter vector β
  - $\hookrightarrow$ ε_i: additive error; i.i.d. with mean 0 and variance σ_ε^2

- We can estimate the model using nonlinear least-squares:

$$\min_\beta \sum_{i=1}^{n} \{y_i - f(\mathbf{x}_i; \beta)\}^2$$

- A nonlinear optimizer is needed to solve for $\hat{\beta}$. More on this when we talk about GMM.

Summary and Readings

- Linear regression
  - ↪ Assumptions of the classical multiple regression model
  - ↪ LS estimator and regression statistics
  - ↪ Multicollinearity
  - ↪ Omitted variables
  - ↪ Dummy variables and nonlinear effects
- Readings
  - ↪ ISL Chapter 3, CLM Chapter 5