Linear Models

Hui Chen

MIT Sloan

15.450, Spring 2023

Outline

Motivation: Modeling Stock Returns

- Market model

$$R^e_{i,t} = \alpha_i + \beta_i R^e_{m,t} + \varepsilon_{i,t}$$

- The Fama-French 3-factor model

$$R^e_{i,t} = \alpha_i + \beta_{i,m}\, R^e_{mt} + \beta_{i,hml}\, HML_t + \beta_{i,smb}\, SMB_t + \varepsilon_{it}$$

- Predicting market returns

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

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

Example: Return Predictability

Cochrane (JF 2011)

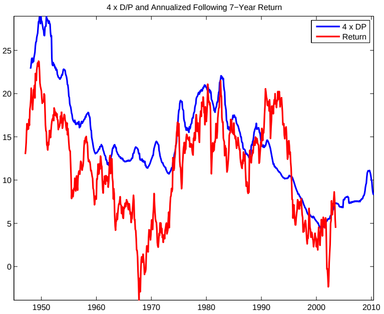


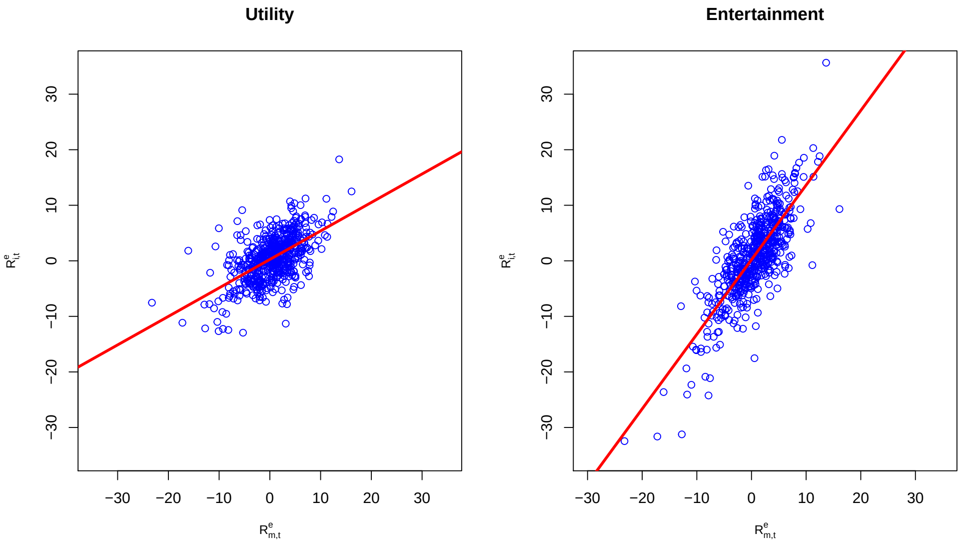
Table I

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



Outline

Simple linear regression

- Univariate linear model (f):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- For any fitted model $\widehat{f}$, with coefficients $(\hat{\beta}_0, \hat{\beta}_1)$, we can compute the fitting errors (residuals):

$$\hat{\varepsilon}_i = y_i - \widehat{y}_i = y_i - \widehat{f}(x_i) = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

- Least squares estimators: Find $(\hat{\beta}_0, \hat{\beta}_1)$ that minimizes the *residual sum of squares* (RSS).

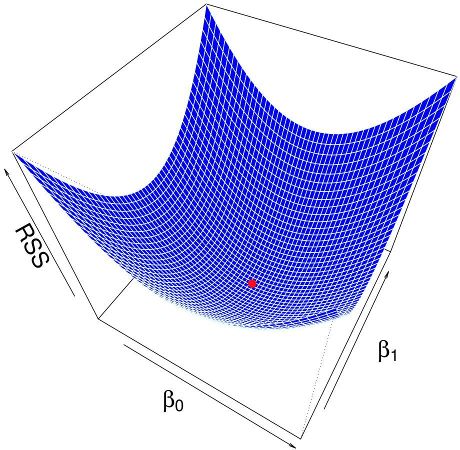
$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

- Solution (more on this later)

Code

```
import statsmodels.formula.api as smf
est = smf.ols('y ~ x', data).fit()
est.summary().tables[1]
```

Least Squares Estimator



Regression Statistics

Estimating market beta for entertainment industry

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  FunRF   R-squared:                       0.667
Model:                            OLS   Adj. R-squared:                  0.666
Method:                 Least Squares   F-statistic:                     2313.
Date:                Thu, 16 Feb 2023   Prob (F-statistic):          4.25e-278
Time:                        00:00:00   Log-Likelihood:                -3585.7
No. Observations:                1158   AIC:                             7175.
Df Residuals:                    1156   BIC:                             7186.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0202      0.159      0.128      0.899      -0.291       0.331
MktRF          1.4147      0.029     48.091      0.000       1.357       1.472
==============================================================================
```

- $\hat{\beta}_1 =$
- $SE(\hat{\beta}_1) =$
- t-statistic $= \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} =$
- p-value
- 95% conf. interval:
- $R^2 = 1 - \frac{RSS}{TSS}$
- Adj. $R^2 =$ $1 - \frac{RSS/(n-p)}{TSS/(n-1)}$
- 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:

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

  ↪ What about the intercept?

- Matrix notation:

$$Y = X\beta + \varepsilon$$

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

- 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: $E[\varepsilon\varepsilon'|X] = \sigma^2\mathbf{I}$

Derivation of Least Squares Estimator

- Find β that minimizes the RSS:

$$\min_{\beta} \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon =$$

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

$$\hat{\beta} =$$

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

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

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

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

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

Outline

1. Introduction
2. Simple linear regression
3. Multiple linear regression
4. **Troubleshooting**
5. Linear regression $\neq$ linear relationship

Independent Predictors

- If the predictor variables are independent, the LS estimates from the multiple linear regression will be the same as obtained by separate simple regressions.
  - $\hookrightarrow$ Run simple regressions with one predictor at a time.
- In such cases, holding σ^2 fixed, more variability in the feature variables reduces the standard errors for $\hat{\beta}$.

Multicollinearity

- Multicollinearity: When two or more predictors are closely related, the accuracy of the least square estimator is substantially reduced.
- To diagnose multicollinearity, compute the variance inflation factor (VIF)

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

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

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

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*.
  - $\hookrightarrow$ Intuition:

- This does not mean we should "overfit" the model by including many features!
  - $\hookrightarrow$ 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.

$$\begin{aligned} b_1 &= (X_1'X_1)^{-1}X_1'Y = (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon) \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon \end{aligned}$$

- 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

- To advise our clients on the design of compensation packages, we want to examine the determinants of CEO compensation across firms.

- 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 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 and/or transform the response variables.
- 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 and Nonlinearity

- With a linear model, we can still capture some nonlinear effects by adding interactions and nonlinear terms (e.g., x_i^2, $\ln x_i$, e^{x_i} ...).
- 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.

Summary and Readings

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