Stat 206: Linear Models

Lecture 12

Nov. 6, 2019

Body Fat: Compare Models

Variables in Model	$\hat{\beta}_1$	$\hat{\beta}_2$	$s\{\hat{\beta}_1\}$	$s\{\hat{\beta}_2\}$	MSE
Model 1: X_1	0.8572	-	0.1288	-	7.95
Model 2: X_2	-	0.8565	-	0.1100	6.3
Model 3: X_1, X_2	0.2224	0.6594	0.3034	0.2912	6.47
Model 4: X_1, X_2, X_3	4.334	-2.857	3.016	2.582	6.15

- The regression coefficient for X_1 (X_2) depending on which other X variables are included in the model.
- The standard errors of the fitted regression coefficients are becoming when more X variables are included into the model.
- MSE tends to as additional X variables are added into the model.

R outputs

- $SSR(X_1) = 352.27$, $SSR(X_1|X_2) = 3.47$.
- The reason why $SSR(X_1|X_2)$ is so small compared to $SSR(X_1)$ is that X_1 and X_2 are with each other and with the response variable Y.
  - When X_2 is already in the model, the marginal contribution from X_1 in explaining Y is since X_2 contains much of the information as X_1 in terms of explaining Y.

What would happen if X_1 and X_2 were not correlated with Y, but were highly correlated among themselves?

In Model 4, none of the three X variables is statistically significant by the T-tests. However, the F-test for regression relation is highly significant. Is there a paradox?

- From the general linear test perspective, each T-test is a test, testing whether the of an X variable is significant given X variables being included in the model.
- The three tests of the marginal effects of X_1, X_2, X_3 together are to testing whether there is a regression relation between Y and (X_1, X_2, X_3).
- The reduced model for each individual test contains X variables and thus may lead to non-significant results due to .
- On the other hand, the reduced model for testing regression relation contains X variable.

Model 4

Effects of Multicollinearity: Summary

- With multicollinearity, the estimated regression coefficients tend to have   sampling variability (i.e., standard errors). This leads to:
  -                    confidence intervals.
  - It's possible that   of the regression coefficients is statistically significant, but at the same time there is a regression relation between the response variable and the entire set of X variables.
- Multicollinearity does not prevent us from getting a   of the data.

prediction

Interpretation of Regression Coefficients and ESS

In the presence of multicollinearity:

- The regression coefficient of an X variable
  which other X variables are also in the model.
- Therefore, a regression coefficient reflect any inherent effect of the corresponding X variable on the response variable, but only a
  given whatever other X variables are also in the model.
- Similarly, there is sum of squares that can be ascribed to any one X variable.
  - The reduction in the total variation in Y ascribed to an X variable must be interpreted as a
    given other X variables also included in the model.

Quantify Multicollinearity: Variance Inflation Factor

Under the standardized model:

$$\sigma^2(\hat{\beta}^*) = \qquad\qquad\qquad\qquad .$$

- The kth diagonal element of the inverse correlation matrix $\mathbf{r}_{XX}^{-1}$ is called the **variance inflation factor (VIF)** for $\hat{\beta}_k^*$, denoted by VIF_k.
- The variance of the estimated regression coefficient $\hat{\beta}_k^*$:

$$\sigma^2(\hat{\beta}_k^*) = \qquad\qquad\qquad\qquad .$$

- The variance of the estimated regression coefficient $\hat{\beta}_k$ in the original model:

$$\sigma^2(\hat{\beta}_k) = \qquad\qquad\qquad\qquad .$$

It can be shown that

$$VIF_k = \frac{1}{1 - R_k^2} (\geq 1), \qquad k = 1, \cdots, p-1,$$

where R_k^2 is the coefficient of multiple determination when X_k is regressed on the rest of X variables $\{X_j : 1 \leq j \neq k \leq p-1\}$.

- If X_k is uncorrelated with the rest of the X variables, then $R_k^2 =$ and $VIF_k =$.
- If $R_k^2 > 0$, then VIF_k , indicating an variance for $\hat{\beta}_k^*$ (eqv. $\hat{\beta}_k$) due to the between X_k and the other X variables.
- If X_k has a perfect linear association with the rest of the X variables, then $R_k^2 =$, $VIF_k =$ and so the variance of $\hat{\beta}_k^*$ (eqv. $\hat{\beta}_k$) is .
- In practice, $\max_k VIF_k > 10$ is often taken as an indication that multicollinearity is high.

Body Fat

Correlation matrices.

$$\mathbf{r}_{XX} = \begin{bmatrix} 1.00 & 0.92 & 0.46 \\ 0.92 & 1.00 & 0.08 \\ 0.46 & 0.08 & 1.00 \end{bmatrix}, \quad \mathbf{r}_{XY} = \begin{bmatrix} 0.84 \\ 0.88 \\ 0.14 \end{bmatrix}.$$

X_1 and X_2 are highly correlated, X_1 and X_3 are moderately correlated, X_2 and X_3 are not much correlated. Moreover,

$$\mathbf{r}_{XX}^{-1} = \begin{bmatrix} 708.84 & -631.92 & -270.99 \\ -631.92 & 564.34 & 241.49 \\ -270.99 & 241.49 & 104.61 \end{bmatrix}$$

So,

Each predictor is with the rest of the predictors.

Coefficient of Partial Determination

It measures the marginal contribution in proportional reduction in SSE by adding one X variable into a model.

- Definition.

$$
\begin{aligned}
& R^2_{Y,j|1,\cdots,j-1,j+1,\cdots,p-1} \\
:= \quad & \frac{SSE(X_1,\cdots,X_{j-1},X_{j+1},\cdots,X_{p-1}) - SSE(X_1,\cdots,X_{p-1})}{SSE(X_1,\cdots,X_{j-1},X_{j+1},\cdots,X_{p-1})} \\
= \quad & \frac{SSR(X_j|X_1,\cdots,X_{j-1},X_{j+1},\cdots,X_{p-1})}{SSE(X_1,\cdots,X_{j-1},X_{j+1},\cdots,X_{p-1})}
\end{aligned}
$$

- Coefficients of partial determination are in between

  .
- For example, $R^2_{Y,1|2} =$ is

Body Fat

- From R outputs, we can obtain a number of coefficients of partial determination. E.g.:

$$R^2_{Y,2|1} =$$

$$R^2_{Y,1|2} = \frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)} = \frac{113.42 - 109.95}{113.43} = 3.1\%.$$

$$R^2_{Y,3|12} =$$

- When X_2 is added to the model containing X_1, SSE is reduced by ; When X_1 is added to the model containing X_2, SSE is reduced by ; When X_3 is added to the model containing X_1, X_2, SSE is reduced by .

R outputs

Interpretation of Coefficient of Partial Determination

- $SSR(X_j|X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_{p-1})$ is the SSR when regressing the residuals $e(Y|X_{-(j)}) = Y - \hat{Y}(X_{-(j)})$ to the residuals $e(X_j|X_{-(j)}) = X_j - \hat{X}_j(X_{-(j)})$, where $X_{-(j)} = \{X_l : 1 \leq l \neq j \leq p\}$.
- So $R^2_{Y,j|1,\cdots,j-1,j+1,\cdots,p-1}$ is the

  between the two sets of residuals obtained by regressing Y and X_j to the rest of variables $X_{-(j)}$, respectively.
- So $R^2_{Y,j|1,\cdots,j-1,j+1,\cdots,p-1}$ measures the linear association between Y and X_j after

  have been adjusted for.

Example. $R^2_{Y,1|2}$.

- Regress Y on X_2: $e_i(Y|X_2) = Y_i - \widehat{Y}_i(X_2), \quad i = 1, \cdots n$.
- Regress X_1 on X_2: $e_i(X_1|X_2) = X_{i1} - \widehat{X}_{i1}(X_2), \quad i = 1, \cdots n$.
- $R^2_{Y1|2}$ equals to the coefficient of simple determination between $e_i(Y|X_2)$ and $e_i(X_1|X_2)$.
- It measures the linear association between Y and X_1 after the linear effects of X_2 have been adjusted for.

Partial Correlations

The **signed** square-root of a coefficient of partial determination is called a partial correlation.

- The sign is the same as the sign of the corresponding fitted regression coefficient.
- Partial correlation is the
  between the .
- Partial correlations can be used to find the “best” X variable to be added next for inclusion in the regression model.

Body Fat

- $r_{Y2|1} =$.
- $r_{Y1|2} =$.
- $r_{Y3|12} =$.

LS Fitted Regression Coefficients as Partial Coefficients

The LS fitted regression coefficients $\hat{\beta}$ are indeed partial coefficients.

- Consider $p-1$ X variables in the model. Let $\hat{\beta}_j$ be the LS fitted regression coefficient for X_j.
- Then $\hat{\beta}_j$ equals to the LS fitted regression coefficient when regressing the residuals $e(Y|X_{-(j)}) = Y - \hat{Y}(X_{-(j)})$ to the residuals $e(X_j|X_{-(j)}) = X_j - \hat{X}_j(X_{-(j)})$, where $X_{-(j)} = \{X_l : 1 \leq l \neq j \leq p\}$.

Confirm this numerically with some of homework data sets.

Polynomial Regression

Polynomial regression models are among the most commonly used models to describe a regression relation.

- Polynomial regression models are very flexible and are easy to fit.
- Polynomial models with higher than third-order terms are rarely employed in practice.
  - They often lead to estimators.
  - They might fit the observed data , but generalize well to new observations, a phenomena called .

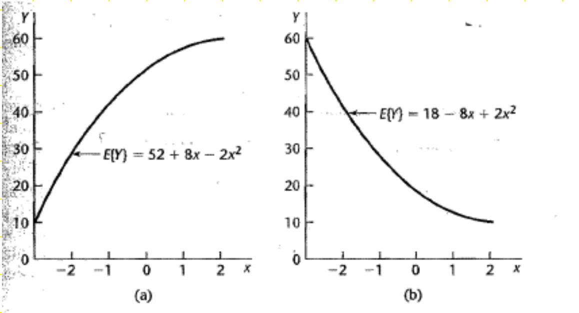
Second-Order Model with One Predictor

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(X_i - \overline{X}) + \beta_2(X_i - \overline{X})^2 + \epsilon_i \\ &= \beta_0 + \beta_1 \tilde{X}_i + \beta_2 \tilde{X}_i^2 + \epsilon_i, \quad i = 1, \cdots, n, \end{aligned}$$

where $\tilde{X}_i = X_i - \overline{X}$ is the centered value of the predictor variable in the ith case.

- Centering often between the linear term X and the quadratic term X^2 substantially *(Why?)* and thus improves numerical accuracy. *Will centering change the fitted regression function?*
- The response function is a parabola:
  - β_0 is the mean response when .
  - β_1 is called the and β_2 is called the .

Figure: Examples of quadratic response functions.



From Applied Linear Statistical Models by Kutner, Nachtsheim, Neter and Li

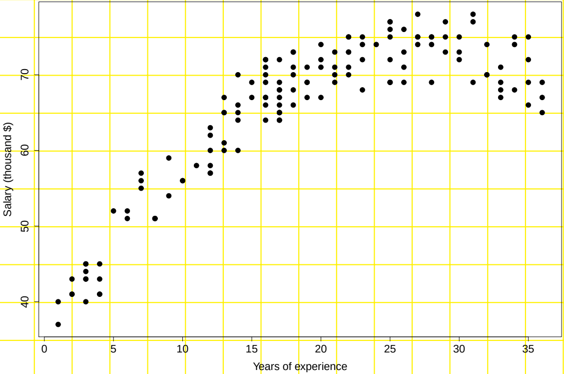
Salary

Professional organizations regularly survey their members for information concerning salaries, pensions, and conditions of employment. One goal is to relate salary to years of experience. This data has years of experience (X) and salary (Y) on 143 cases. [1]

Case	Salary($) Y	Experience(Years) X
1	71	26
2	69	19
3	73	22
...	...	...
141	67	16
142	71	20
143	69	31

[1] Source of data: Tryfos (1998): Methods for business analysis and forecasting

Figure: Scatter plot of salary versus years of experience



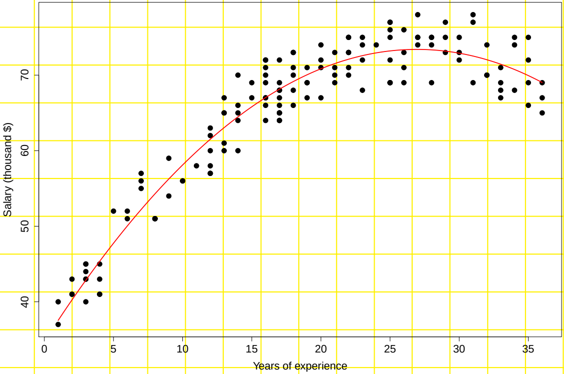
The regression relation between salary and years of experience appears to be curvilinear.

Salary: Second-Order Model

```
> salary.c=salary
> salary.c[,"Experience"]=salary[,"Experience"]-mean(salary[,"Experience"]) ## center the X variable
> fitc=lm(Salary~ Experience+I(Experience^2), data=salary.c) ## fit a second-order model
> summary(fitc)

Call:
lm(formula = Salary ~ Experience + I(Experience^2), data = salary.c)

Residuals:
Min       1Q  Median      3Q     Max
-4.5786 -2.3573  0.0957  2.0171  5.5176

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)      69.927208   0.323090  216.43   <2e-16 ***
Experience        0.861177   0.024957   34.51   <2e-16 ***
I(Experience^2) -0.053316   0.002477  -21.53   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.817 on 140 degrees of freedom
Multiple R-squared: 0.9247,     Adjusted R-squared: 0.9236
F-statistic: 859.3 on 2 and 140 DF,  p-value: < 2.2e-16
```

Figure: Fitted response function:

$$y = 69.93 + 0.861 \times (X - 18.86) - 0.0533 \times (X - 18.86)^2$$



Second-Order Model with Two Predictors

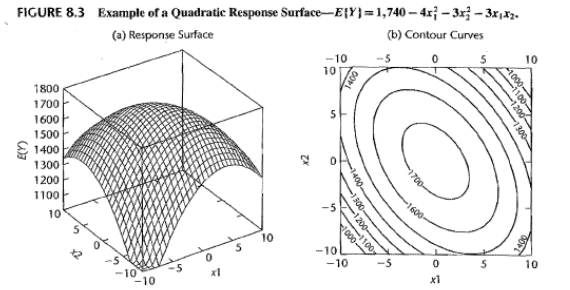
where $\tilde{X}_{i1} = X_{i1} - \overline{X}_1$, $\tilde{X}_{i2} = X_{i2} - \overline{X}_2$.

- Response function is a conic section:

$$E(Y) = \beta_0 + \beta_1 \tilde{X}_1 + \beta_2 \tilde{X}_2 + \beta_{11} \tilde{X}_1^2 + \beta_{22} \tilde{X}_2^2 + \beta_{12} \tilde{X}_1 \tilde{X}_2.$$

- This model contains separate and terms for each of the two predictors.
- It also contains a term representing the between the two predictors.
- β_{12} is called the .

Figure: A quadratic response surface.



From Applied Linear Statistical Models by Kutner, Nachtsheim, Neter and Li

The contour curves show various combinations of the values of the two predictors that yield the same value of the response function.

Second-Order Model with K Predictors

$$Y_i = \beta_0 + \sum_{k=1}^{K} \beta_k \tilde{X}_{ik} + \sum_{k=1}^{K} \beta_{kk} \tilde{X}_{ik}^2 + \sum_{1 \le k < k' \le K} \beta_{kk'} \tilde{X}_{ik} \tilde{X}_{ik'} + \epsilon_i, i = 1, \cdots, n,$$

where $\tilde{X}_{ik} = X_{ik} - \overline{X}_k$ $(k = 1, \cdots, K)$.

- Response function:

$$E(Y) = \beta_0 + \sum_{k=1}^{K} \beta_k \tilde{X}_k + \sum_{k=1}^{K} \beta_{kk} \tilde{X}_k^2 + \sum_{1 \le k < k' \le K} \beta_{kk'} \tilde{X}_k \tilde{X}_{k'}.$$

- β_ks are linear effect coefficients; β_{kk}s are quadratic effect coefficients.
- $\{\beta_{kk'} : 1 \le k < k' \le K\}$ are interaction effect coefficients between respective pairs of predictors. (The cross-product terms are second-order terms.)

Salary: Third-Order Model

```
> fit3=lm(Salary~ Experience+I(Experience^2)+I(Experience^3), data=salary.c)
> summary(fit3)
Call:
lm(formula = Salary ~ Experience + I(Experience^2) + I(Experience^3),
data = salary.c)
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)      69.9484745  0.3224575  216.92   <2e-16 ***
Experience        0.9364986  0.0603531   15.52   <2e-16 ***
I(Experience^2) -0.0537196  0.0024866  -21.60   <2e-16 ***
I(Experience^3) -0.0003957  0.0002888   -1.37    0.173
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.808 on 139 degrees of freedom
Multiple R-squared: 0.9257,     Adjusted R-squared: 0.9241
F-statistic: 577.1 on 3 and 139 DF,  p-value: < 2.2e-16

> anova(fit3)
Analysis of Variance Table
Response: Salary
Df Sum Sq Mean Sq   F value Pr(>F)
Experience         1 9962.9   9962.9 1263.1043 <2e-16 ***
I(Experience^2)    1 3677.9   3677.9  466.2810 <2e-16 ***
I(Experience^3)    1   14.8     14.8    1.8764  0.173
Residuals        139 1096.4      7.9
```

- First test whether the third-order term may be dropped.
  - Full model: third-order model vs. reduced model: second-order model.
  - $SSR(X^3|X, X^2) = 14.8$ with d.f. 1 and $SSE(X, X^2, X^3) = 1096.4$ with d.f. 139. The F-statistic is 1.876 and pvalue is 0.173.
  - Therefore, the third-order term is not significant and may be dropped.
- Then test whether the second-order term may be dropped.
  - Full model: second-order model vs. reduced model: first-order model.
  - $SSR(X^2|X) = 3677.9$ with d.f. 1 and $SSE(X, X^2) = SSE(X, X^2, X^3) + SSR(X^3|X, X^2) = 1111.2$ with d.f. 140. The F-statistic is 466.28 and pvalue $< 2e-16$.
  - So the second-order term is significant and should be retained.
- Thus the first-order term should also be retained and we end up with the second-order model.