ch03_notes

August 8, 2019

0.1 # Linear Regression

0.2 Simple Linear Regression

- For data (X, Y), $X, Y \in \mathbb{R}$, ***simple linear regression*** models Y as a linear function of X

$$Y = \beta_0 + \beta_1 X + \epsilon$$

and predicts

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

where $\hat{\beta}_i$ is the estimate for β_i.

0.2.1 Estimating the Coefficients

Estimates of the coefficients β_0, β_1 arize from minimizing ***residual sum of squares***

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

using calculus one finds estimates7

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

These are sometimes called the ***least squares estimates***.

0.2.2 Assessing the Accuracy of the Coefficent Estimates

- The ***population regression line***8 is the line given by

$$Y = \beta_0 + \beta_1 X$$

and the ***least squares regression line*** is the line given by

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X$$

- The least squares estimate is an unbiased estimator 9

- Assuming errors ϵ_i are uncorrelated with common variance $\sigma^2 = \mathbb{V}(\epsilon)$, the standard errors of $\hat{\beta}_0, \hat{\beta}_1$ are

$$\mathbf{se}(\hat{\beta}_0) = \sigma\sqrt{\left[\frac{1}{n} + \frac{\overline{x}}{\sum_i (x_i - \overline{x})^2}\right]}$$

$$\mathbf{se}(\hat{\beta}_1) = \sigma\sqrt{\frac{1}{\sum_i (x_i - \overline{x})^2}}$$

- The estimated standard errors $\hat{\mathbf{se}}(\hat{\beta}_0), \hat{\mathbf{se}}(\hat{\beta}_0)$ are found by estimating σ with the ***residual standard error*** 10

$$\hat{\sigma} = RSE := \sqrt{\frac{RSS}{n-2}}$$

- Approximate $1 - \alpha$ ***confidence intervals*** 11 for the least squares estimators are

$$\hat{\beta}_i \pm t_{\alpha/2}\hat{\mathbf{se}}(\hat{\beta}_i)$$

- Most common hypothesis tests for the least squares estimates are

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

the rejection region is

$$\{x \in \mathbb{R} \mid T > t\}$$

where t is the test-statistic 12

$$t = \frac{\hat{\beta}_i - \beta_i}{\hat{\mathbf{se}}(\hat{\beta}_i)}$$

0.2.3 Assessing the Accuracy of the Model

Quality of fit (model accuracy) is commonly assessed using RSE and the R^2 statistic.

0.2.4 Residual Standard Errors

- The RSE is a measure of the overall difference between the observed responses y_i and the predicted responses $\hat{y}_i$. Thus it provides a measure of *lack-of-fit* of the model – higher RSE indicates worse fit.
- RSE is measured in units of Y so it provides an absolute measure of lack of fit, which is sometimes difficult to interpret

0.2.5 R^2 Statistic

- The R^2 statistic is

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

where $TSS = \sum_i (y_i - \overline{y})^2$ is the ***total sum of squares***.

- TSS measures the total variability in Y, while RSS measures the variability left after modeling Y by $f(X)$. Thus, R^2 measures the proportion of variability in Y that can be explained by the model. R^2 is dimensionless so it provides a good relative measure of lack-of-fit.

- As $R^2 \to 1$, the model explains more of the variability in Y. As $R^2 \to 0$, the model explains less 13. What constitutes a good R^2 value depends on context

- We can also think of R^2 as a measure of the linear relationship between Y and X. Another such measure is the correlation $\text{corr}(X, Y)$, which is estimated by the sample correlation r. In the case of simple linear regression, $R^2 = r^2$.

0.3 Multiple Linear Regression

- For data (X, Y), $X = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$,$Y \in \mathbb{R}$, ***multiple linear regression*** models Y as a linear function 14 of X

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

and predicts

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p + \epsilon$$

where $\hat{\beta}_i$ is the estimate of β_i

- If we form the $n \times (p+1)$ matrix $\mathbf{X}$ with rows $(1, X_{i1}, \ldots, X_{ip})$, response vector $Y = (Y_1, \ldots, Y_n)$, parameter vector $\beta = (\beta_0, \ldots, \beta_p)$ and noise vector $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ then the model can be written in matrix form

$$Y = \mathbf{X}\beta + \epsilon$$

0.3.1 Estimating the Regression Coefficients

- RSS is defined and estimates $\hat{\beta}_i$ for the parameters β_i are chosen to minimize RSS 15 as in the Section 0.2.1.

- If the data matrix $\mathbf{X}$ has full rank, then the estimate 16 $\hat{\beta}$ for the parameter vector is

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \beta$$

0.3.2 Important Questions

Is There a Relationship Between the Response and Predictors?

- One way to answer this question is a hypothesis test

$$\begin{aligned} H_0 &: \beta_i = 0 && \text{for all } 1 \leqslant i \leqslant p \\ H_a &: \beta_i \neq 0 && \text{for some } 1 \leqslant i \leqslant p \end{aligned}$$

- The test statistic is the ***F-statistic***17

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

  where TSS, RSS are defined as in simple linear regression.

- Assuming the model is correct,

$$\mathbb{E}\left(\frac{RSS}{n-p-1}\right) = \sigma^2$$

  where again, $\sigma^2 = \mathbb{V}(\epsilon)$. Further assuming H_0 is true,

$$\mathbb{E}\left(\frac{TSS - TSS}{p}\right) = \sigma^2$$

  hence $H_0 \Rightarrow F \approx 1$ and $H_a \Rightarrow F > 118$.

- Another way to answer this question is a hypothesis test on a subset of the predictors of size q

$$\begin{aligned} H_0 &: \beta_i = 0 && \text{for all } p-q+1 \leqslant i \leqslant p \\ H_a &: \beta_i \neq 0 && \text{for some } p-q+1 \leqslant i \leqslant p \end{aligned}$$

  where RSS_0 is the residual sum of squares for a second model ommitting the last q predictors. The F-statistic is

$$F = \frac{(RSS_0 - RSS)/p}{RSS/(n-p-1)}$$

- These hypothesis tests help us conclude that at least one of the predictors is related to the response (the second test narrows it down a bit), but don't indicate which ones.

Deciding on Important Variables

- The task of finding which predictors are related to the response is sometimes known as ***variable selection***.19
- Various statistics can be used to judge the quality of models using different subsets of the predictors. Examples are ***Mallows*** C_p ***criterion***, ***Akaike Information Criterion (AIC)***, ***Bayesian Information Criterion*** and ***adjusted*** R^2.
- Since the number of distinct linear regression models grows exponentially with p exhaustive search is infeasible unless p is small. Common approaches to consider a smaller set of possible models are
  - ***Forward Selection*** Start with ***the null model*** M_0 (an intercept but no predictors). Fit p simple regressions and add to the null model the one with lowest RSS, resulting in a new model M_1. Iterate until a stopping rule is reached.
  - ***Backward Selection*** Start with a model M_p consisting of all predictors. Remove the variable with largest p-value, resulting in a new model M_{p-1}. Iterate until a stopping rule is reached.
  - ***Mixed Selection*** Proceed with forward selection, but remove any predictors whose p-value is too large.

Model Fit

- As in simple regression, RSE and R^2 are two common measures of model fit
- In multiple regression, $R^2 = Corr(Y, \hat{Y})^2$, with the same interpretation as in simple regression. The model $\hat{Y}$ maximizes R^2 among all linear models.
- R^2 increases monotonically in the number of predictors, but small increases indicate the low relative value of the corresponding predictor.
- In multiple regression

$$RSS = \sqrt{\frac{RSS}{n-p-1}}$$

- Visualization can be helpful in assessing model fit, e.g. by suggesting the inclusion of ***interaction*** terms

Predictions There are 3 types of uncertainty associated with predicting Y by $\hat{Y}$

- ***Estimation Error***. $\hat{Y} = \hat{f}(X)$ is only an estimate $f(X)$. This error is reducible. We can compute confidence intervals to quantify it.
- ***Model Bias***. A linear form for $f(X)$ may be inappropriate. This error is also reducible
- ***Noise***. The noise term ϵ is a random variable. This error is irreducible. We can compute ***predeiction intervals*** to quantify it.

0.4 Other Considerations In the Regression Model

0.4.1 Qualitative Predictors

- If the i-th predictor X_i is a factor (qualitative) with K ***levels*** (that is K possible values) then we model it by $K-1$ indicator variables (sometimes called a ***dummy variables***).
- Two commons definitions of the dummy variables are

$$\tilde{X}_i = \begin{cases} 1 & X_i = k \\ 0 & X_i \neq k \end{cases}$$

$$\tilde{X}_i = \begin{cases} 1 & X_i = k \\ -1 & X_i \neq k \end{cases}$$

  for $1 \leqslant k \leqslant K$.

- The corresponding regression model is

$$Y = \beta_0 + \sum_i \beta_i \tilde{X}_i + \epsilon$$

  since we can only have $\tilde{X}_i = 1$ if $\tilde{X}_j \neq 1$ for $j \neq i$, this model can be seen as K distinct models

$$Y = \begin{cases} \beta_0 & X_i = 1 \\ \beta_0 + \beta_1 & X_i = 2 \\ \vdots & \vdots \\ \beta_0 + \beta_K & X_i = K \end{cases}$$

0.4.2 Extensions of the Linear Model

The standard linear regression we have been discussing relies on the twin assumptions

- ***Additivity***: The effect of X_i on Y is independent of the effect of X_j for $j \neq i$.
- ***Linearity***: Y is linear in X_i for all i.

We can extend the model by relaxing these assumptions

Removing the Additive Assumption

- Dropping the assumption of additivity leads to the possible inclusion of ***interaction*** or ***synergy*** effects among predictors.
- One way to model an interaction effect between predictors X_i and X_j is to include an ***interaction term***, $\beta_{i+j} X_i X_j$. The non-interaction terms $\beta_i X_i$ model the ***main effects***.
- We can perform hypothesis tests as in the standard linear model to select important terms/variables. However, the ***hierarchical principle*** dictates that, if we include an interaction effect, we should include the corresponding main effects, even if the latter aren't statistically significant.

Non-linear Relationships

- Dropping the assumption of linearity leads to the possible includion of non-linear effects.
- One common way to model non-linearity is to use ***polynomial regression*** 20, that is model $f(X)$ with a polynomial in the predictors. For example in the case of a single predictor X

$$Y = \beta_0 + \beta_1 X + \cdots + \beta_d X^s$$

  models Y as a degree d polynomial in X
- In general one can model a non-linear effect of predictors X_i by including a non-linear function of the X_i in the model

0.4.3 Potential Problems

Non-linearity of the Data ***Residual plots*** are a useful way of vizualizing non-linearity. The presence of a discernible pattern may indicate a problem with the linearity of the model.

Correlation of Error Terms

- Standard linear regression assumes $\text{Corr}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.
- Correlated error terms frequently occur in the context of ***time series***.
- Positively correlated error terms may display ***tracking*** behavior (adjacent residuals may have similar values).

Non-constant Variance of Error Terms

- Standard linear regression assumes the variance of errors is constant across observations, i.e. $\mathbb{V}(\epsilon_i) = \sigma^2$ for all $1 \leqslant i \leqslant n$
- ***Hetereoscedasticity***, or variance which changes across observations can be identified by a funnel shape in the residual plot.
- One way to reduce hetereoscedasticity is to transform Y by a concave function such as $\log Y$ or $\sqrt{Y}$.
- Another way to do this is ***weighted least squares***. This weights terms in RSS with weights w_i inversely proportional to σ_i^2 where $\sigma_i^2 = \mathbb{V}(\epsilon_i)$.

Outliers

- An ***outlier*** is an observation for which the value of y_i given x_i is unusual, i.e. such that the squared-error $(y_i - \hat{y}_i)^2$ is large
- Outliers can have disproportionate effects on statistics e.g. R^2, which in turn affect the entire analysis (e.g. confidence intervals, hypothesis tests).
- Residual plots can identify outliers. In practice, we plot ***studentized residuals***

$$\frac{\hat{\epsilon}_i}{\hat{\mathbf{se}}(\hat{\epsilon}_i)}$$

- If an outlier is due to a data collection error it can be removed, but great care should be taken when doing this.

High Leverage Points

- A ***high leverage point*** is a point with an unusual value of x_i.
- High leverage points tend to have a sizable impact on $\hat{f}$.
- To quantify the leverage of x_i, we use the ***leverage statistic***. In simple linear regression this is

$$h_i = \frac{1}{n} + \frac{(X_j - \overline{X})^2}{\sum_j (X_j - \overline{X})^2}$$

Collinearity

- ***Collinearity*** is a linear relationship among two or more predictors.
- Collinearity reduces the accuracy of coefficient estimates 21
- Collinearity reduces the ***power***22 of the hypothesis test
- Collinearity between two variables can be detected by the sample correlation matrix $\hat{\Sigma}$. A high value for
$$|(\hat{\Sigma})_{ij}| = |\text{corr}(\hat{X}_i, X_j)|$$
indicates high correlation between X_i, X_j hence high collinearity in the data23.
- ***Multicollinearity*** is a linear relationship among more than two predictors.
- Multicollinearity can be detected using the ***variance inflation factor*** (VIF)24.
$$VIF(\hat{\beta}_i) = \frac{1}{1 - R^2_{X_i|X_{-i}}}$$
where $R^2_{X_i|X_{-i}}$ is the R^2 from regression of X_i onto all other predictors.
- One solution to the presence of collinearity is to drop one of the problematic variables, which is usually not an issue, since correlation among variables is seen as redundant.
- Another solution is to combine the problematic variables into a single predictor (e.g. an average)

0.5 The Marketing Plan

Skip

0.6 Comparison of Linear Regression and K-Nearest Neighbors

- Linear regression is a parametric model for regression (with parameter $\beta = (\beta_0, \ldots, \beta_p)$).
- KNN regression is a popular non-parametric model, which estimates

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i$$

- In general, a parametric model will outperform a non-parametric model if the parametric estimation $\hat{f}$ is close to the true f.
- KNN regression suffers from the ***curse of dimensionality*** - as the dimension increases the data become sparse. Effectively this is a reduction in sample size, hence KNN performance commonly decreases as the dimension p increases.
- In general parametric methods outperform non-parametric methods when there is a small number of observations per predictor.
- Even if performance of KNN and linear regression is comparable, the latter may be favored for interpretability.

0.7 Footnotes

7. The value $(\hat{\beta}_0, \hat{\beta}_1)$ is the local minimum in $\mathbb{R}^2$ of the "loss function" given by RSS
8. Here estimate means the same as "estimator", found elsewhere in the statistics literature. The population regression line is given by the "true" (population) values (β_0, β_1) of the parameter, while the least squares line is given by the estimator $(\hat{\beta}_0, \hat{\beta}_1)$
9. In other words, $\mathbb{E}\left((\hat{\beta}_0, \hat{\beta}_1)\right) = (\beta_0, \beta_1)$
10. The factor $\frac{1}{n-2}$ is a correction to make this an unbiased estimator, the quantity $n-2$ is known as the "degrees of freedom". Note this is a special case of $n-p-1$ degrees of freedom for p predictors where $p = 1$.
11. This appears to be based on the assumption (no doubt proved in the literature) that the least squares estimators are asymptotically t-distributed, $\hat{\beta}_i \approx Student_{n-2}(\beta_i, \hat{\mathbf{se}}(\hat{\beta}_i))$.
12. This is the Wald test for the statistic T, which (by footnote 4) has $T \approx Student_{n-2}(0, 1)$.
13. This can happen if either the model is wrong (i.e. a linear form for $f(X)$ isn't a good choice) or because $\mathbb{V}(\epsilon)$ is large.
14. This estimation method is known as ***Ordinary Least Squares (OLS)***. The estimate is the solution to the quadratic minimzation problem

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \, ||y - \mathbf{X}\beta||^2$$

15. The estimate is any solution to the quadratic minimzation problem

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} ||y - \mathbf{X}\beta||^2$$

which can be found by solving the normal equations

$$\mathbf{X}^\top \mathbf{X}\hat{\beta} = \mathbf{X}^\top y$$

16. If $\mathbf{X}$ has full rank then $\mathbf{X}^\top \mathbf{X}$ is invertible and the normal equations have a unique solution

17. Assuming the ϵ_i are normally distributed, $\epsilon_i \sim N(\mu_i, \sigma^2)$ where $\mu = \beta_0 + \sum \beta_i X_i$), the F-statistic has an F-distribution with $p, n - p$ degrees of freedom (F has this asymptotic distribution even without the normality assumption).

The use of the F statistic arises from ANOVA among the predictors, which is beyond our scope. There is some qualitative discussion of the motivation for the F statistic on page 77 of the text. It is an appropriate statistic in the case p i

18. How much $F > 1$ should be before we rejct H_0 depends on n and p. If n is large, F need not be much greater than 1, and if it's small,

19. This is discussed extensively in chapter 6.

20. This is discussed in chapter 7.

21. This is due to issues identifying the global minimum of RSS. In the example in the text, in the presence of collinearity, the global minimum is in a long "valley". The coefficient estimates are very sensitive to the data – small changes in the data yeild large changes in the estimates.

22. The power of the test is the probability of correctly rejecting $H_0 : \beta_i = 0$, i.e. correctly accepting $H_a : \beta_i \neq 0$. Since it uncreases uncertainty of the coefficient estimates, it increases $\hat{se}(\hat{\beta}_i)$, hence reduces the t-statistic, making it less likely H_0 is rejected.

23. However, the converse is not true – absence of such entries in the sample correlation matrix doesn't indicate absence of collinearity. The matrix only detects pairwise correlation, and a predictor may correlate two or more other predictors.

24. This is defined the ratio of the (sample) variance of $\hat{\beta}_i$ when fitting the full model divided by the variance of $\hat{\beta}_i$ when fit on it's own. It can be computed using the given formula.
