ELEMENTS OF STATISTICAL LEARNING - CHAPTER SOLUTIONS

ANDREW TULLOCH

3. Chapter 3

Exercise 3.1. *Show that the F statistic for dropping a single coefficient from a model is equal to the square of the corresponding z-score.*

Proof. Recall that the F statistic is defined by the following expression

$$\frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)}.$$

where RSS_0, RSS_1 and $p_0 + 1, p_1 + 1$ refer to the residual sum of squares and the number of free parameters in the smaller and bigger models, respectively. Recall also that the F statistic has a $F_{p_1 - p_0, N - p_1 - 1}$ distribution under the null hypothesis that the smaller model is correct.

Next, recall that the z-score of a coefficient is

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$$

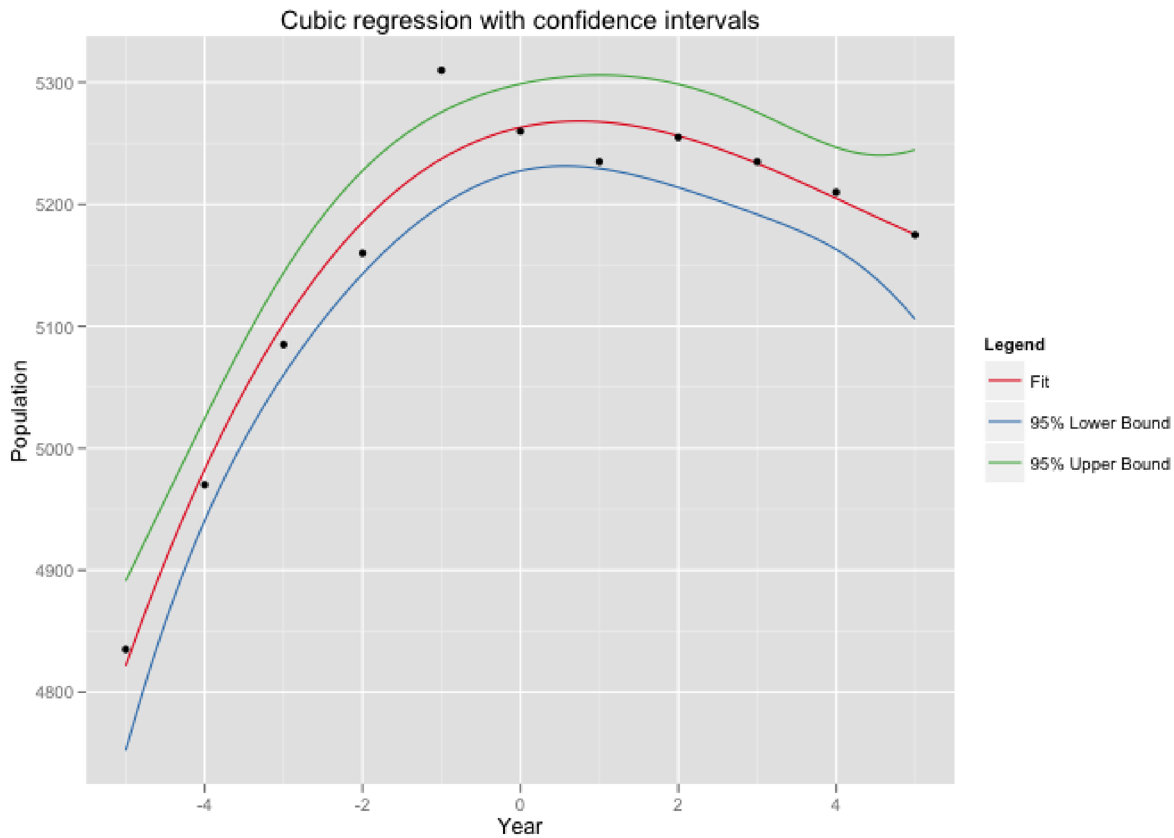
and under the null hypothesis that β_j is zero, z_j is distributed according to a t-distribution with $N - p - 1$ degrees of freedom.

Hence, by dropping a single coefficient from a model, our F statistic has a $F_{1, N-p-1}$ where $p + 1$ are the number of parameters in the original model. Similarly, the corresponding z-score is distributed according to a t_{N-p-1} distribution, and thus the square of the z-score is distributed according to an $F_{1,N-p-1}$ distribution, as required. $\square$

Exercise 3.2. *Given data on two variables X and Y, consider fitting a cubic polynomial regression model $f(X) = \sum_{j=0}^{3} \beta_j X^j$. In addition to plotting the fitted curve, you would like a 95% confidence band about the curve. Consider the following two approaches:*

1. *At each point x_0, form a 95% confidence interval for the linear function $a^T\beta = \sum_{j=0}^{3} \beta_j x_0^j$.*
2. *Form a 95% confidence set for β as in (3.15), which in tun generates confidence intervals for $f(x_0)$.*

How do these approaches differ? Which band is likely to be wider? Conduct a small simulation experiment to compare the two methods.



Proof. The key distinction is that in the first case, we form the set of points such that we are 95% confident that $\hat{f}(x_0)$ is within this set, whereas in the second method, we are 95% confident that an arbitrary point is within our confidence interval. This is the distinction between a *pointwise* approach and a *global* confidence estimate.

In the pointwise approach, we seek to estimate the variance of an individual prediction - that is, to calculate $\mathrm{Var}(\hat{f}(x_0)|x_0)$. Here, we have

$$\begin{aligned}\sigma_0^2 = \mathrm{Var}(\hat{f}(x_0)|x_0) &= \mathrm{Var}(x_0^T\hat{\beta}|x_0)\\ &= x_0^T\mathrm{Var}(\hat{\beta})x_0\\ &= \hat{\sigma}^2 x_0^T(X^TX)^{-1}x_0.\end{aligned}$$

where $\hat{\sigma}^2$ is the estimated variance of the innovations ϵ_i.

We can implement this algorithm in R as follows:

```
library("ggplot2")
library("reshape2")

# Raw data
simulation.xs <- c(1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, ↩
    1968, 1969)
simulation.ys <- c(4835, 4970, 5085, 5160, 5310, 5260, 5235, 5255, 5235, ↩
    5210, 5175)
simulation.df <- data.frame(pop = simulation.ys, year = simulation.xs)

# Rescale years
simulation.df$year <- simulation.df$year - 1964

# Generate regression, construct confidence intervals
fit <- lm(pop ~ year + I(year^2) + I(year^3), data=simulation.df)
xs <- seq(-5, 5, 0.1)
fit.confidence <- predict(fit, data.frame(year=xs), interval="confidence", ↩
    level=0.95)


# Create data frame containing variables of interest
df <- as.data.frame(fit.confidence)
df$year <- xs
df = melt(df, id.vars="year")

p <- ggplot()
p <- p + geom_line(aes(x=year, y=value, colour=variable),
                   df)
P <- p + geom_point(aes(x=year, y=pop),
                    simulation.df)
p <- p + scale_x_continuous('Year')
p <- p + scale_y_continuous('Population')
p <- p + opts(title="Cubic regression with confidence intervals")
p <- p + scale_color_brewer(name="Legend",
        labels=c("Fit",
                 "95% Lower Bound",
                 "95% Upper Bound"),
        palette="Set1")
```

□

Exercise 3.3. *Prove the Gauss-Markov theorem: the least squares estimate of a parameter $a^T\beta$ has a variance no bigger than that of any other linear unbiased estimate of $a^T\beta$.*

Secondly, show that if $\hat{V}$ is the variance-covariance matrix of the least squares estimate of β and $\tilde{V}$ is the variance covariance matrix of any other linear unbiased estimate, then $\hat{V} \leq \tilde{V}$, where $B \leq A$ if $A - B$ is positive semidefinite.

Proof. Let $\hat{\theta} = a^T\hat{\beta} = a^T(X^TX)^{-1}X^Ty$ be the least squares estimate of $a^T\beta$. Let $\tilde{\theta} = c^Ty$ be any other unbiased linear estimator of $a^T\beta$. Now, let $d^T = c^T - a^T(X^{-1}X)^{-1}X^T$. Then as c^Ty is unbiased, we must have

$$\begin{aligned} E(c^Ty) &= E\left(a^T(X^TX)^{-1}X^T + d^T\right)y \\ &= a^T\beta + d^TX\beta \\ &= a^T\beta \end{aligned}$$

as c^Ty is unbiased, which implies that $d^TX = 0$.

Now we calculate the variance of our estimator. We have

$$\begin{aligned} \mathrm{Var}(c^Ty) &= c^T\mathrm{Var}(y)c \\ &= \sigma^2c^Tc \\ &= \sigma^2\left(a^T(X^TX)^{-1}X^T + d^T\right)\left(a^T(X^TX)^{-1}X^T + d^T\right)^T \\ &= \sigma^2\left(a^T(X^TX)^{-1}X^T + d^T\right)\left(X(X^TX)^{-1}a + d\right) \\ &= \sigma^2\left(a^T(X^TX)^{-1}X^TX(X^TX)^{-1}a + a^T(X^TX)^{-1}\underbrace{X^Td}_{=0} + \underbrace{d^TX}_{=0}(X^TX)^{-1}a + d^Td\right) \\ &= \sigma^2\left(\underbrace{a^T(X^TX)^{-1}a}_{\mathrm{Var}(\hat{\theta})} + \underbrace{d^td}_{\geq 0}\right) \end{aligned}$$

Thus $\mathrm{Var}(\hat{\theta}) \leq \mathrm{Var}(\tilde{\theta})$ for all other unbiased linear estimators $\tilde{\theta}$.

The proof of the matrix version is almost identical, except we replace our vector d with a matrix D. It is then possible to show that $\tilde{V} = \hat{V} + D^TD$, and as D^TD is a positive semidefinite matrix for any D, we have $\hat{V} \leq \tilde{V}$. □

Exercise 3.4. *Show how the vector of least square coefficients can be obtained from a single pass of the Gram-Schmidt procedure. Represent your solution in terms of the QR decomposition of X.*

Proof. Recall that by a single pass of the Gram-Schmidt procedure, we can write our matrix X as

$$X = Z\Gamma,$$

where Z contains the orthogonal columns z_j, and Γ is an upper-diagonal matrix with ones on the diagonal, and $\gamma_{ij} = \frac{\langle z_i, x_j \rangle}{\|z_i\|^2}$. This is a reflection of the fact that by definition,

$$x_j = z_j + \sum_{k=0}^{j-1} \gamma_{kj} z_k.$$

Now, by the QR decomposition, we can write $X = QR$, where Q is an orthogonal matrix and R is an upper triangular matrix. We have $Q = ZD^{-1}$ and $R = D\Gamma$, where D is a diagonal matrix with $D_{jj} = \|z_j\|$.

Now, by definition of $\hat{\beta}$, we have

$$(X^T X)\hat{\beta} = X^T y.$$

Now, using the QR decomposition, we have

$$\begin{aligned}
(R^T Q^T)(QR)\hat{\beta} &= R^T Q^T y \\
R\hat{\beta} &= Q^T y
\end{aligned}$$

As R is upper triangular, we can write

$$\begin{aligned}
R_{pp}\hat{\beta}_p &= \langle q_p, y \rangle \\
\|z_p\|\hat{\beta}_p &= \|z_p\|^{-1}\langle z_p, y \rangle \\
\hat{\beta}_p &= \frac{\langle z_p, y \rangle}{\|z_p\|^2}
\end{aligned}$$

in accordance with our previous results. Now, by back substitution, we can obtain the sequence of regression coefficients $\hat{\beta}_j$. As an example, to calculate $\hat{\beta}_{p-1}$, we have

$$\begin{aligned}
R_{p-1,p-1}\hat{\beta}_{p-1} + R_{p-1,p}\hat{\beta}_p &= \langle q_{p-1}, y \rangle \\
\|z_{p-1}\|\hat{\beta}_{p-1} + \|z_{p-1}\|\gamma_{p-1,p}\hat{\beta}_p &= \|z_{p-1}\|^{-1}\langle z_{p-1}, y \rangle
\end{aligned}$$

and then solving for $\hat{\beta}_{p-1}$. This process can be repeated for all β_j, thus obtaining the regression coefficients in one pass of the Gram-Schmidt procedure. □

Exercise 3.5. *Consider the ridge regression problem (3.41). Show that this problem is equivalent to the problem*

$$\hat{\beta}^c = \operatorname*{arg\,min}_{\beta^c} \left(\sum_{i=1}^N \left(y_i - \beta_0^c - \sum_{j=1}^p (x_{ij} - \hat{x}_j)\beta_j^c \right)^2 + \lambda \sum_{j=1}^p \beta_j^{c2} \right)^2.$$

Proof. Consider rewriting our objective function above as

$$L(\beta^c) = \sum_{i=1}^N \left(y_i - \left(\beta_0^c - \sum_{j=1}^p \bar{x}_j \beta_j^c \right) - \sum_{j=1}^p x_{ij}\beta_j^c \right)^2 + \lambda \sum_{j=1}^p \beta_j^{2\,2}$$

Note that making the substitutions

$$\beta_0 \mapsto \beta_0^c - \sum_{j=1}^{p} \hat{x}_j \beta_j$$

$$\beta_j \mapsto \beta_j^c, j = 1, 2, \ldots, p$$

that $\hat{\beta}$ is a minimiser of the original ridge regression equation if $\hat{\beta}^c$ is a minimiser of our modified ridge regression.

The modified solution merely has a shifted intercept term, and all other coefficients remain the same. $\square$