36-401 Modern Regression Exam #1 Solutions

October 5, 2017

Problem 1 [40 pts.]

(a) In Lecture Notes 4 we derived the following estimators for the simple linear regression model:

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X} \tag{1}$$

$$\widehat{\beta}_1 = \frac{c_{XY}}{s_X^2}, \tag{2}$$

where

$$c_{XY} = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})(Y_j - \overline{Y}) \quad \text{and} \quad s_X^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2.$$

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{c_{XY}}{s_X^2} \\
&= \frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})Y_j - \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})\overline{Y}}{s_X^2} \\
&= \frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})(\beta_0 + \beta_1 X_j + \epsilon_j) - \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})(\beta_0 + \beta_1\overline{X} + \overline{\epsilon})}{s_X^2} \\
&= \frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})(\beta_1 X_j - \beta_1\overline{X} + \epsilon_j - \overline{\epsilon})}{s_X^2} \\
&= \beta_1 \cdot \frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2}{s_X^2} + \sum_{j=1}^{n}\left(\frac{X_j - \overline{X}}{n s_X^2}\right)(\epsilon_j - \overline{\epsilon}) \\
&= \beta_1 + \sum_{j=1}^{n}\left(\frac{X_j - \overline{X}}{n s_X^2}\right)(\epsilon_j - \overline{\epsilon}) \\
&= \beta_1 + \frac{1}{n s_X^2}\sum_{j=1}^{n}\left(X_j\epsilon_j - \overline{\epsilon}X_j - \overline{X}\epsilon_j + \overline{X}\,\overline{\epsilon}\right) \\
&= \beta_1 + \frac{1}{n s_X^2}\sum_{j=1}^{n}\left(X_j\epsilon_j - \overline{X}\epsilon_j\right) \\
&= \beta_1 + \frac{1}{n s_X^2}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)\epsilon_j
\end{aligned}$$

(a)

$$\begin{aligned}
\widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X} \\
&= \overline{Y} - \left(\beta_1 + \frac{1}{n s_X^2} \sum_{j=1}^{n} \left(X_j - \overline{X}\right)\epsilon_j\right)\overline{X} \\
&= \beta_0 + \beta_1 \overline{X} + \overline{\epsilon} - \left(\beta_1 + \frac{1}{n s_X^2} \sum_{j=1}^{n} \left(X_j - \overline{X}\right)\epsilon_j\right)\overline{X} \\
&= \beta_0 + \overline{\epsilon} - \frac{1}{n s_X^2} \sum_{j=1}^{n} \overline{X}\left(X_j - \overline{X}\right)\epsilon_j \\
&= \beta_0 + \frac{1}{n}\sum_{j=1}^{n} \epsilon_j - \sum_{j=1}^{n} \overline{X} \frac{X_j - \overline{X}}{n s_X^2}\epsilon_j \\
&= \beta_0 + \sum_{j=1}^{n} \left(\frac{1}{n} - \overline{X}\frac{X_j - \overline{X}}{n s_X^2}\right)\epsilon_j
\end{aligned}$$

(b)

$$\begin{aligned}
\widehat{m}(x) &= \widehat{\beta}_0 + \widehat{\beta}_1 x \\
&= \left[\beta_0 + \sum_{j=1}^{n} \left(\frac{1}{n} - \overline{X}\frac{X_j - \overline{X}}{n s_X^2}\right)\epsilon_j\right] + \left[\beta_1 + \frac{1}{n s_X^2} \sum_{j=1}^{n} \left(X_j - \overline{X}\right)\epsilon_j\right] \cdot x \\
&= \left[\beta_0 + \beta_1 x\right] + \left[\sum_{j=1}^{n} \left(\frac{1}{n} - \overline{X}\frac{X_j - \overline{X}}{n s_X^2}\right)\epsilon_j + \frac{1}{n s_X^2} \sum_{j=1}^{n} \left(X_j - \overline{X}\right)\epsilon_j \cdot x\right] \\
&= m(x) + \left[\sum_{j=1}^{n} \left(\frac{1}{n} - \overline{X}\frac{X_j - \overline{X}}{n s_X^2}\right)\epsilon_j + \sum_{j=1}^{n} \frac{X_j - \overline{X}}{n s_X^2}\epsilon_j \cdot x\right] \\
&= m(x) + \left[\sum_{j=1}^{n} \left(\frac{1}{n} - \overline{X}\frac{X_j - \overline{X}}{n s_X^2}\right)\epsilon_j + \frac{X_j - \overline{X}}{n s_X^2}\epsilon_j \cdot x\right] \\
&= m(x) + \sum_{j=1}^{n} \left(\frac{1}{n} + (x - \overline{X})\frac{X_j - \overline{X}}{n s_X^2}\right)\epsilon_j
\end{aligned}$$

(c)

$$\begin{aligned}
e_i &= Y_i - \widehat{m}(X_i) \\
&= \beta_0 + \beta_1 X_i + \epsilon_i - \left[m(X_i) + \sum_{j=1}^{n} \left(\frac{1}{n} + (X_i - \overline{X})\frac{X_j - \overline{X}}{n s_X^2}\right)\epsilon_j\right] \\
&= m(X_i) + \epsilon_i - \left[m(X_i) + \sum_{j=1}^{n} \left(\frac{1}{n} + (X_i - \overline{X})\frac{X_j - \overline{X}}{n s_X^2}\right)\epsilon_j\right] \\
&= \epsilon_i - \sum_{j=1}^{n} \left(\frac{1}{n} + (X_i - \overline{X})\frac{X_j - \overline{X}}{n s_X^2}\right)\epsilon_j \\
&= \sum_{j=1}^{n} \left(\delta_{ij} - \frac{1}{n} - (X_i - \overline{X})\frac{X_j - \overline{X}}{n s_X^2}\right)\epsilon_j
\end{aligned}$$

(d) Let

$$c_{ij} = \left(\delta_{ij} - \frac{1}{n} - (X_i - \overline{X})\frac{X_j - \overline{X}}{ns_X^2}\right).$$

Then

$$\begin{aligned}
\mathbb{E}[e_i^2] &= \mathbb{E}\left[\left(\sum_{j=1}^n c_{ij}\epsilon_j\right)^2\right] \\
&= \mathrm{Var}\left(\sum_{j=1}^n c_{ij}\epsilon_j\right) + \left(\mathbb{E}\left[\sum_{j=1}^n c_{ij}\epsilon_j\right]\right)^2 \\
&= \mathrm{Var}\left(\sum_{j=1}^n c_{ij}\epsilon_j\right) + \left(\sum_{j=1}^n c_{ij}\underbrace{\mathbb{E}[\epsilon_j]}_{=0}\right)^2 \\
&= \mathrm{Var}\left(\sum_{j=1}^n c_{ij}\epsilon_j\right) \\
&= \mathrm{Cov}\left(\sum_{j=1}^n c_{ij}\epsilon_j,\ \sum_{k=1}^n c_{ik}\epsilon_k\right) \\
&= \sum_{j=1}^n c_{ij}\sum_{k=1}^n c_{ik}\cdot\underbrace{\mathrm{Cov}\big(\epsilon_j,\ \epsilon_k\big)}_{=0 \text{ when } j\neq k} \\
&= \sum_{j=1}^n c_{ij}^2\mathrm{Cov}\big(\epsilon_j,\ \epsilon_j\big) \\
&= \sum_{j=1}^n c_{ij}^2\mathrm{Var}\big(\epsilon_j\big) \\
&= \sigma^2\sum_{j=1}^n c_{ij}^2 \\
&= \sigma^2\sum_{j=1}^n\left(\delta_{ij} - \frac{1}{n} - (X_i - \overline{X})\frac{X_j - \overline{X}}{ns_X^2}\right)^2 \\
&= \sigma^2\sum_{j=1}^n\left(\delta_{ij}^2 - \frac{2}{n}\delta_{ij} - 2\delta_{ij}(X_i - \overline{X})\frac{X_j - \overline{X}}{ns_X^2} + \frac{1}{n^2} + \frac{2(X_i - \overline{X})}{n}\frac{X_j - \overline{X}}{ns_X^2} + (X_i - \overline{X})^2\frac{(X_j - \overline{X})^2}{n^2s_X^4}\right) \\
&= \sigma^2\left(1 - \frac{1}{n} - 2\frac{(X_i - \overline{X})^2}{ns_X^2}\right) + \sigma^2\sum_{j=1}^n\left(\frac{2(X_i - \overline{X})(X_j - \overline{X})}{n^2s_X^2} + \frac{(X_i - \overline{X})^2(X_j - \overline{X})^2}{n^2s_X^4}\right) \\
&= \sigma^2\left(1 - \frac{1}{n} - 2\frac{(X_i - \overline{X})^2}{ns_X^2}\right) + \frac{2(X_i - \overline{X})\sigma^2}{n^2s_X^2}\underbrace{\sum_{j=1}^n(X_j - \overline{X})}_{=0} + \frac{(X_i - \overline{X})^2\sigma^2}{n^2s_X^4}\underbrace{\sum_{j=1}^n(X_j - \overline{X})^2}_{=n\cdot s_X^2} \\
&= \sigma^2\left(1 - \frac{1}{n} - 2\frac{(X_i - \overline{X})^2}{ns_X^2}\right) + \frac{(X_i - \overline{X})^2\sigma^2}{ns_X^2} \\
&= \sigma^2\left(1 - \frac{1}{n} - \frac{(X_i - \overline{X})^2}{ns_X^2}\right).
\end{aligned}$$

(e)

$$\begin{aligned}
\mathbb{E}[\widehat{\sigma}^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} e_j^2\right] \\
&= \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[e_j^2] \\
&= \frac{1}{n}\sum_{j=1}^{n} \sigma^2\left(1 - \frac{1}{n} - \frac{(X_j - \overline{X})^2}{n s_X^2}\right) \\
&= \sigma^2 - \frac{\sigma^2}{n} - \frac{\sigma^2}{n s_X^2} \cdot \sum_{j=1}^{n} \frac{(X_j - \overline{X})^2}{n} \\
&= \sigma^2 - \frac{\sigma^2}{n} - \frac{\sigma^2}{n} \\
&= \frac{n-2}{n}\sigma^2
\end{aligned}$$

Problem 2 [20 pts.]

(a) Using the fact that $SSE/\sigma^2 \sim \chi^2_{n-2}$, we have

$$P\left(\chi^2_{n-2}(\alpha/2) < \frac{SSE}{\sigma^2} < \chi^2_{n-2}(1-\alpha/2)\right) = 1-\alpha$$

$$P\left(\frac{1}{\chi^2_{n-2}(1-\alpha/2)} < \frac{\sigma^2}{SSE} < \frac{1}{\chi^2_{n-2}(\alpha/2)}\right) = 1-\alpha$$

$$P\left(\frac{SSE}{\chi^2_{n-2}(1-\alpha/2)} < \sigma^2 < \frac{SSE}{\chi^2_{n-2}(\alpha/2)}\right) = 1-\alpha.$$

Thus,

$$\left(\frac{SSE}{\chi^2_{n-2}(1-\alpha/2)}, \frac{SSE}{\chi^2_{n-2}(\alpha/2)}\right)$$

is a $1-\alpha$ confidence interval for σ^2.

(b) Plugging these values into the interval we derived in part (a) gives

$$\left(\frac{100}{\chi^2_{41}(0.975)}, \frac{100}{\chi^2_{41}(0.025)}\right)$$
$$= \left(\frac{100}{60.6}, \frac{100}{25.2}\right)$$
$$\approx \Big(1.65,\ 3.97\Big).$$

(c) No, we cannot reject the hypothesis $H_0 : \sigma^2 = 3$ at level α because this value lies within the $1-\alpha$ confidence interval for σ^2, implying the test statistic does not lie in the rejection region.

Problem 3 [40 pts.]

(a) The covariate is `population` and the response is `pcgmp` (per-capita gross metropolitan product). A $\log_{10}$ transformation was applied to `population`.

(b)
$$\widehat{m}(\log_{10}(\texttt{pop})) = -23306 + 10246 \cdot \log_{10}(\texttt{pop})$$

(c)
$$\begin{aligned}\widehat{m}(\log_{10}(10^6)) &= -23306 + 10246 \cdot 6 \\ &= 38170.\end{aligned}$$

$$\begin{aligned}\widehat{m}(\log_{10}(200000)) &= -23306 + 10246 \cdot \log_{10}(200000) \\ &= -23306 + 10246 \cdot \log_{10}(2 \cdot 10^5) \\ &= -23306 + 10246 \cdot \log_{10}(2) + 10246 \cdot \log_{10}(10^5) \\ &= -23306 + 10246 \cdot 0.301 + 10246 \cdot 5 \\ &= 31008.05\end{aligned}$$

Yes, these numbers are in keeping with the trend that `pcgmp` tends to increase, on the average, as `population` increases.

(d) Yes.
$$\begin{aligned}\widehat{m}(0) &= -23306 + 10246 \cdot 0 \\ &= -23306.\end{aligned}$$

Note, however, that such an estimate does not make sense in the context of the problem. This is because it is an extrapolation to the average `pcgmp` for a city with `pop` $= 1$.

(e)
$$P\left(\widehat{\beta}_1 - t_{n-2}(1-\alpha/2) \cdot \widehat{\text{se}}(\widehat{\beta}_1) < \beta_1 < \widehat{\beta}_1 + t_{n-2}(1-\alpha/2) \cdot \widehat{\text{se}}(\widehat{\beta}_1)\right) = 1 - \alpha$$
$$P\left(\widehat{\beta}_1 - t_{364}(0.975) \cdot \widehat{\text{se}}(\widehat{\beta}_1) < \beta_1 < \widehat{\beta}_1 + t_{364}(0.975) \cdot \widehat{\text{se}}(\widehat{\beta}_1)\right) = 0.95$$
$$P\left(10246 - 1.97 \cdot 900 < \beta_1 < 10246 + 1.97 \cdot 900\right) = 0.95$$
$$P\big(8473 < \beta_1 < 12019\big) = 0.95$$

Therefore, a 95% confidence interval for β_1 is
$$\big(8473,\ 12019\big).$$

(f)
$$\begin{aligned}7930 &= \texttt{Residual standard error} \\ &= \sqrt{\frac{1}{364}\sum_{j=1}^{366} e_i^2}\end{aligned}$$

Therefore,

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{366}\sum_{j=1}^{366} e_i^2 \\ &= \frac{364}{366}\cdot 7930^2 \\ &= 62541267\end{aligned}$$

(g) Although a bit tedious, the information in the summary output can be used to compute

$$s_X^2 = \frac{1}{n}\sum_{j=1}^{n}\left(\log_{10}(\texttt{pop}_j) - \overline{\log_{10}(\texttt{pop})}\right)^2.$$

However, from this, it is still not possible to transform back to the sample variance on the original scale

$$s_{\texttt{pop}}^2 = \frac{1}{n}\sum_{j=1}^{n}\left(\texttt{pop}_j - \overline{\texttt{pop}}\right)^2.$$

(h) Nothing in the output tests the assumption that the underlying mean function is actually linear. For this we must examine residual plots.