STAT210A - Homework 10

Hoang Duong

November 24, 2014

Problem 1. Read Chapter of Keener

Problem 2. Complete Sufficient Statistics

Proof. We have the density for Y is:

$$
\begin{aligned}
p_Y(y) =& \frac{1}{(2\pi)^{-n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)\right\} && (1)\\
=& \frac{1}{(2\pi)^{-n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\left(y-X\hat{\beta}+X\hat{\beta}-X\beta\right)^T\left(y-X\hat{\beta}+X\hat{\beta}-X\beta\right)\right\} && (2)\\
=& \frac{1}{(2\pi)^{-n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\left[\|y-X\hat{\beta}\|_2^2+\|X\hat{\beta}-X\beta\|_2^2+2\left(y-X\hat{\beta}\right)^T\left(X\hat{\beta}-X\beta\right)\right]\right\} && (3)\\
=& \frac{1}{(2\pi)^{-n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\left[(n-p)S^2+\|X\hat{\beta}-X\beta\|_2^2+2y^T(I_n-H)^TX(\hat{\beta}-\beta)\right]\right\} && (4)
\end{aligned}
$$

For $H = X\left(X^TX\right)^{-1}X^T$ is the projection matrix. H is symmetric, thus $I_n - X$ is symmetric. So $(I_n-H)^T X = (I_n-H)X = X - HX = 0$. So the cross term in (4) is zero. Thus:

$$p_Y(y) == \frac{1}{(2\pi)^{-n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\left[(n-p)S^2+\|X\hat{\beta}-X\beta\|_2^2\right]\right\}$$

By the property of exponential family, we have $\left(\hat{\beta}, S^2\right)$ is a complete sufficient statistics. □

Problem 3. Hypothesis Testing

Proof. (a) Consider a test $\delta_b(X) = \mathbb{I}_{\left\{|\hat{\beta}|\geq b\right\}}$ of rejecting H_0 if $\left|\hat{\beta}\right| \geq b$, and fail to reject if not. In the setting of Keener Section 14.5, we have:

$$
\begin{aligned}
S^2 =& \frac{1}{n-2}\sum e^2\\
\mathbb{P}\left[\hat{\beta}_2 - \frac{St_{\alpha/2,n-2}}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}} \leq \beta \leq \hat{\beta}_2 + \frac{St_{\alpha/2,n-2}}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}\right] =& 1-\alpha\\
\mathbb{P}\left[\left|\hat{\beta}-\beta\right| \geq \frac{St_{\alpha/2,n-2}}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}\right] =& \alpha
\end{aligned}
$$

So if we choose $b = \frac{St_{\alpha/2,n-2}}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}$, we have $\delta_b(X)$ is a level$-\alpha$ test for $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$.

(b) Idea taken from John P. Buonaccorsi's note at UMass, and Parker (2011).

Consider $\hat{\theta} = \bar{x} + \left(y_0 - \hat{\beta}_1\right)/\hat{\beta}_2$ as an estimate for θ. From Casella and Berger (2002), we note:

$$
\begin{aligned}
\operatorname{Var}\left[\frac{U}{V}\right] \approx& \frac{\operatorname{Var}[U]}{\mathbb{E}^2 V} + \frac{\mathbb{E}^2 U}{\mathbb{E}^4 V}\operatorname{Var}[V] - 2\frac{\mathbb{E}U}{\mathbb{E}^3 V}\operatorname{Cov}[U,V] \\
U =& y_0 - \hat{\beta}_1 \\
V =& \hat{\beta}_2 \\
\mathbb{E}U =& \mathbb{E}\left[y_0 - \hat{\beta}_1\right] = \mathbb{E}\left[\hat{\beta}_2(\theta - \bar{x}) + e_0\right] \\
=& (\theta - \bar{x})\,\mathbb{E}\hat{\beta}_2 = \beta_2(\theta - \bar{x}) \\
\mathbb{E}V =& \beta_2 \\
\operatorname{Var}U =& \operatorname{Var}[y_0] + \operatorname{Var}\left[\hat{\beta}_1\right] - 2\operatorname{Cov}\left[y_0, \hat{\beta}_1\right] \\
=& \sigma^2 + \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
\operatorname{Var}V =& \frac{\sigma^2}{S_{xx}} \\
\operatorname{Cov}[U,V] =& \frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
\Rightarrow \operatorname{Var}\left[\frac{U}{V}\right] \approx& \frac{\sigma^2}{\beta_2^2}\left(1 + \frac{1}{n} + \frac{(\theta - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)
\end{aligned}
$$

Based on the Delta Method, we have:

$$
\mathbb{P}\left[\hat{\theta} - t_{1-\alpha/2,n-2}\frac{\hat{\sigma}^2}{\hat{\beta}_2^2}\sqrt{1 + \frac{1}{n} + \frac{(\theta - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \le \theta \le \hat{\theta} + t_{1-\alpha/2,n-2}\frac{\hat{\sigma}^2}{\hat{\beta}_2^2}\sqrt{1 + \frac{1}{n} + \frac{(\theta - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right] = 1 - \alpha
$$

Thus a level-α test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ is to reject H_0 iff:

$$
\begin{aligned}
\left|\hat{\theta} - \theta_0\right| \ge& t_{1-\alpha/2,n-2}\frac{\hat{\sigma}^2}{\hat{\beta}_2^2}\sqrt{1 + \frac{1}{n} + \frac{(\theta_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \\
\hat{\theta} =& \bar{x} + \frac{y_0 - \hat{\beta}_1}{\hat{\beta}_2} \\
\hat{\sigma}^2 =& \frac{1}{n-2}\sum_{i=1}^n \left(y_i - \hat{\beta}_1 - \hat{\beta}_2(x_i - \bar{x})\right)^2
\end{aligned}
$$

(c) We have:

$$
h(\theta) = \frac{y_0 - \left(\hat{\beta}_1 + \hat{\beta}_2\theta\right)}{\left[\sigma^2 + \frac{\sigma^2}{n} + \theta^2\frac{\sigma^2}{S_{xx}} + 2\theta\frac{\sigma^2\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]^{1/2}}
$$

is t-distributed with $n-2$ degrees of freedom. The set of θ where $h^2(\theta) \le t^2_{1-\alpha/2,n-2}$ is a $1-\alpha$ confidence region for θ. □

Problem 4. Confidence Interval

Proof. Rewrite the regression as a linear model:

$$\begin{aligned}
\begin{bmatrix} Y_1 \\ \dots \\ Y_{n_1} \\ Y_{n_1+1} \\ \dots \\ Y_n \end{bmatrix} &= \begin{bmatrix} 1 & x_1 & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 1 & x_{n_1} & 0 & 0 \\ 0 & 0 & 1 & x_{n_1+1} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \dots \\ \epsilon_{n_1} \\ \epsilon_{n_1+1} \\ \dots \\ \epsilon_n \end{bmatrix} \\
Y &= X\beta + \epsilon \\
\epsilon &\sim \mathcal{N}\left(0, \sigma^2 I\right)
\end{aligned}$$

The OLS estimator for this linear model is:

$$\begin{aligned}
\hat{\beta} &= X\left(X^T X\right)^{-1} X^T y \\
\hat{y} &= X\hat{\beta} \\
\text{Var}\hat{\beta} &= \sigma^2 \left(X^T X\right)^{-1} \\
c &= \begin{bmatrix} 0 \\ -1 \\ 0 \\ 1 \end{bmatrix} \\
\mathbb{E}\left[c^T \hat{\beta}\right] &= c^T \beta = \beta_4 - \beta_2 \\
\Rightarrow \text{Var}\left[c^T \hat{\beta}\right] &= \sigma^2 c^T \left(X^T X\right)^{-1} c \\
\hat{\sigma}^2 &= \frac{1}{n-4} \|y - \hat{y}\|_2^2 \\
\Rightarrow \text{SE}\left(c^T \hat{\beta}\right) &= \hat{\sigma}^2 c^T \left(X^T X\right)^{-1} c
\end{aligned}$$

We have $\left(c^T \hat{\beta} - c^T \beta\right) / SE(c^T \beta)$ follows a standard t distrubution with $n-4$ degree of freedom. Thus:

$$\begin{aligned}
\mathbb{P}\left[-t_{\alpha/2,n-4} \leq \frac{c^T \hat{\beta} - c^T \beta}{SE(c^T \hat{\beta})} \leq t_{\alpha/2,n-4}\right] &= 1 - \alpha \\
\Rightarrow \mathbb{P}\left[-c^T \hat{\beta} - t_{\alpha/2,n-4} SE(c^T \hat{\beta}) \leq -c^T \beta \leq -c^T \hat{\beta} + t_{\alpha/2,n-4} SE\left(c^T \hat{\beta}\right)\right] &= 1 - \alpha \\
\Rightarrow \mathbb{P}\left[c^T \hat{\beta} - t_{\alpha/2,n-4} SE\left(c^T \hat{\beta}\right) \leq c^T \beta \leq c^T \hat{\beta} + t_{\alpha/2,n-4} SE(c^T \hat{\beta})\right] &= 1 - \alpha
\end{aligned}$$

So the $(1-\alpha)$ confident interval for $\beta_4 - \beta_2$ is:

$$\begin{aligned}
&\left[c^T \hat{\beta} - t_{\alpha/2,n-4} SE(c^T \hat{\beta}), c^T \hat{\beta} + t_{\alpha/2,n-4} SE(c^T \hat{\beta})\right] \\
c^T &= [0, -1, 0, 1] \\
\hat{\beta} &= \left(X^T X\right)^{-1} X^T y \\
SE\left(c^T \hat{\beta}\right) &= \hat{\sigma}^2 c^T \left(X^T X\right)^{-1} c \\
\hat{\sigma}^2 &= \frac{1}{n-4} \|y - \hat{y}\|_2^2 \\
\hat{y} &= X\hat{\beta}
\end{aligned}$$

$t_{\alpha/2,n-4}$ is the the value such that the CDF of standard t-distribution with $n-4$ degree of freedom evaluate to $1 - \alpha/2$. $\square$

Problem 5. Log-normal Distribution

Proof. (a) Let $X = \log Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$

$$\begin{aligned}
\mathbb{E}Y =& \mathbb{E}\exp X \\
=& \int \exp\left\{x\right\} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
=& \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{x^2 - 2x\mu + \mu^2 - 2x\sigma^2}{2\sigma^2}\right\} dx \\
=& \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\mu + \frac{\sigma^2}{2}\right\} \exp\left\{-\frac{x^2 - 2x(\mu+\sigma^2) + \mu^2 + 2\mu\sigma^2 + \sigma^4}{2\sigma^2}\right\} dx \\
=& \exp\left\{\mu + \frac{\sigma^2}{2}\right\} \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{x^2 - 2x(\mu+\sigma^2) + \mu^2 + 2\mu\sigma^2 + \sigma^4}{2\sigma^2}\right\} dx \\
=& \exp\left\{\mu + \frac{\sigma^2}{2}\right\} \\
\mathbb{E}Y^2 =& \mathbb{E}\exp 2X \\
=& \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{x^2 - 2x\mu + \mu^2 - 4x\sigma^2}{2\sigma^2}\right\} dx \\
=& \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{2\mu + 2\sigma^2\right\} \exp\left\{-\frac{x^2 - 2x(\mu+2\sigma^2) + \mu^2 + 4\mu\sigma^2 + 4\sigma^4}{2\sigma^2}\right\} dx \\
=& \exp\left\{2\mu + 2\sigma^2\right\} \\
\mathbb{E}Y^\alpha =& \exp\left\{\alpha\mu + \alpha^2\sigma^2/2\right\} \\
\Rightarrow \mathrm{Var}Y =& \mathbb{E}Y^2 - \mathbb{E}^2 Y = \exp\left\{2\mu + 2\sigma^2\right\} - \exp\left\{2\mu + \sigma^2\right\} \\
\mathbb{P}\left[Y \le y\right] =& \mathbb{P}\left[\log Y \le \log y\right] = \Phi\left(\log y\right) \\
\Rightarrow p_Y(y) =& \frac{\partial \Phi(\log y)}{\partial y} \\
=& \frac{\partial \Phi\left(\log y\right)}{\partial \log y} \frac{\partial \log y}{\partial y} \\
=& \phi(\log y)\frac{1}{y} \\
=& \frac{1}{\sqrt{2\pi}\sigma y} \exp\left\{-\frac{(\log y - \mu)^2}{2\sigma^2}\right\}
\end{aligned}$$

(b) $Y_1, ..., Y_n$ are i.i.d log-normal is equivalent to $\log Y_1, \log Y_2, ..., \log Y_n$ are i.i.d normal μ, σ^2. In the Gaussian setting, we know that sample mean and sample variance is the UMVU. Thus $\left(\sum_{i=1}^n \log Y_i\right)/n$ is the UMVU for μ.

(c) Consider $\hat{\eta}_1 = \exp\left\{\left(\sum_{i=1}^n \log Y_i\right)/n\right\}$ as an estimate for $\eta = \mathbb{E}Y_i$. We have:

$$
\begin{aligned}
\mathbb{E}\hat{\eta}_1 =& \mathbb{E}\left[\prod_{i=1}^{n} Y_i^{1/n}\right] \\
=& \prod_{i=1}^{n} \mathbb{E}\left[Y_i^{1/n}\right] \quad Y_i\text{'s are independent} \\
=& \prod_{i=1}^{n} \exp\left\{\frac{\mu}{n} + \frac{\sigma^2}{2n^2}\right\} \\
=& \exp\left\{\mu + \frac{\sigma^2}{2n}\right\} \\
=& \eta \exp\left\{\frac{\sigma^2}{2n}\right\} \\
\Rightarrow \mathbb{E}\frac{\hat{\eta}_1}{\exp\{\sigma^2/(2n)\}} =& \eta \\
\Leftrightarrow \mathbb{E}\left[\exp\left\{\frac{1}{n}\sum_{i=1}^{n}\log Y_i + \frac{1}{2n}\sigma^2\right\}\right] =& \eta \\
\hat{\eta} =& \exp\left\{\frac{1}{n}\sum_{i=1}^{n}\log Y_i + \frac{1}{2n}\sigma^2\right\}
\end{aligned}
$$

So $\hat{\eta}$ is an unbiased estimator of η. Now consider the density of $(Y_1, Y_2, ..., Y_n)$:

$$
\begin{aligned}
p_{Y_1,...,Y_n}(y_1,...,y_n) =& \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}\sigma y_i}\exp\left\{-\frac{(\log y_i - \mu)^2}{2\sigma^2}\right\}\right] \\
=& \frac{1}{(2\pi)^{n/2}\sigma^n \prod_{i=1}^{n} y_i}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}\log^2 y_i - 2\mu\sum\log y_i + n\mu^2\right]\right\}
\end{aligned}
$$

With σ^2 a known constant, by the factor theorem for sufficient statistics, we have $\sum \log y_i$ is a sufficient statistic. $\hat{\eta}$ is a function of sufficient statistic, thus it is UMVU.

(d) I would log-transform Y_i into $\log Y_i$ then perform typical OLS on $\log Y_i \sim \beta_1 + \beta_2 x_i$. So the estimator should be $\left(X^T X\right)^{-1} X^T \log Y$. For $\log Y$ is the element-wise log of $Y = [Y_1, ..., Y_n]$. This estimator is also the MLE estimator according to the multivariate log-normal density we have in (c).

References:

Buonaccorsi, J.P., 2012. STAT505/ST697R: Regression Analysis, Fall 2012 Note. http://people.math.umass.edu/˜johnpb/

Casella, G., Berger, R.L. (2002). Statistical Inference, 2nd ed., Duxbury, CA.

Parker, P.A., Vining, G.G., Wilson, S.R., Szarka III, J.L., Johnson, N.G., 2011. The Prediction Properties of Inverse and Reverse Regression for the Simple Linear Calibration Problem. http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa

□