MVE 190 / MSG 500 Exam 2013-12-19 Solutions

Question 1

a) $Q(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2$

$$Q'(\beta) = 2\sum_{i=1}^{n} (y_i - \beta x_i)(-x_i) = 2\sum_{i=1}^{n} -x_i y_i + 2\beta \sum_{i=1}^{n} x_i^2$$

$$Q'(\beta) = 0 \iff -\sum_{i=1}^{n} x_i y_i + \beta \sum_{i=1}^{n} x_i^2 = 0$$

$$\iff \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

b) $$\mathrm{var}(\hat{\beta}) = \mathrm{var}\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \right) = \frac{\mathrm{var}\left(\sum_{i=1}^{n} x_i y_i\right)}{\left(\sum_{i=1}^{n} x_i^2\right)^2} = \frac{\sum_{i=1}^{n} \mathrm{var}(x_i y_i)}{\left(\sum_{i=1}^{n} x_i^2\right)^2}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 \,\mathrm{var}(y_i)}{\left(\sum_{i=1}^{n} x_i^2\right)^2} = \left\{ \text{assume } \varepsilon_i \sim N(0, \sigma^2) \right\} = \frac{\sum_{i=1}^{n} x_i^2 \sigma^2}{\left(\sum_{i=1}^{n} |x_i^2|\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

$$\mathrm{var}(\hat{y}_i) = \mathrm{var}(\hat{\beta} x_i) = x_i^2 \,\mathrm{var}(\hat{\beta}) = \frac{x_i^2 \sigma^2}{\sum_{i=1}^{n} x_i^2}$$

c) $$E[\hat{\beta}_1] = E\left[\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \right] = \frac{\sum_{i=1}^{n} E[y_i]}{\sum_{i=1}^{n} x_i} = \frac{\sum_{i=1}^{n} \beta x_i}{\sum_{i=1}^{n} x_i} = \beta \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i} = \beta$$

$$E[\hat{\beta}_2] = E\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \right] = \frac{\sum_{i=1}^{n} x_i E[y_i]}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i \beta x_i}{\sum_{i=1}^{n} x_i^2} = \beta \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} = \beta$$

We have $\text{var}(\hat{\beta}_2)$ from b)

$$\text{var}(\hat{\beta}_1) = \text{var}\left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}\right) = \frac{\sum_{i=1}^{n} \text{var}(y_i)}{\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} \sigma^2}{\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{n\sigma^2}{\left(\sum_{i=1}^{n} x_i\right)^2}$$

we have $\text{var}(\hat{\beta}_1) = \frac{n\sigma^2}{\left(\sum_{i=1}^{n} x_i\right)^2}$ and $\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$

dividing by $n\sigma^2$ we get $\frac{1}{\left(\sum_{i=1}^{n} x_i\right)^2}$, $\frac{1}{n\sum_{i=1}^{n} x_i^2}$

in general, if $x_i > 0$ we have that $n\sum_{i=1}^{n} x_i^2 > \left(\sum_{i=1}^{n} x_i\right)^2$

which implies $\text{var}(\hat{\beta}_2) < \text{var}(\hat{\beta}_1)$

e) $\hat{\beta}_2$ is the slope that minimizes the perpendicular distances of the observations to the regression line (definition of least-squares) while $\hat{\beta}_1$ is just the average slope of the observations.

$\hat{\beta}_2$ is a better estimate in the sense that it has a smaller variance.

Question 2

a) Let X = # times a word appear. We assume $X \sim Poisson(\lambda)$

The mle of λ is $\hat{\lambda} = \bar{x} = \frac{1 \cdot 16432 + 2 \cdot 4776 + \cdots + 10 \cdot 222}{16432 + 4776 + \cdots + 222} = 2.08$

Thus $Pr(X = k) = \frac{2.08^k e^{-2.08}}{k!}$

The predicted counts are given by (Total # words) $\times$ $Pr(X = k)$

Predicted counts
7187
7461
5163
2679
1112
385
114
29
6
1

which doesn't seem to be an adequate model

b) Model 1 is the simple poisson model. Corresponds to the predicted counts computed above.

Model 2 is the non-transformed GLM and Model 3 is the transformed one. Figures 2 and 3 correspond to models 2 and 3 respectively since the diagnostics in figure 3 are better than those in figure 2.

Question 3

a) Let $\hat{\beta}_1 = \sum_{i=1}^{n} k_i y_i$, where $k_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

$$Var(\hat{\beta}_1) = \sum_{i=1}^{n} var(k_i y_i) = \sum_{i=1}^{n} k_i^2 var(y_i) = \sum_{i=1}^{n} k_i^2 var(\varepsilon_i)$$

$$= \frac{(1-\delta)\sigma^2 + \delta\tau^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

b) Small perturbations of the data means small δ. In that case the $var(\hat{\beta}_1)$ is almost the usual variance (the one for non-perturbed data), so we expect the estimator to perform well.

c) Model 1 and 2 are the OLS and LAD for non-perturbed data since, in that case, both are expected to perform equally good. Model 3 is the OLS for perturbe data: since it's not a robust method, it is strongly driven by the oulliers. Model 4 is the LAD for the perturbed data, since it's robust to outliers is expected to perform better than OLS and be closer to the models for non perturbed data.

Question 4

a) No, linear regression doesn't imply causality

b) The men working in that company are, in average, more experienced than women

Question 5

a) Assume the counts for cell i,j come from a multinomial distribution: $n_{i,j}$: count for cel i,j →

$(n_{11}, n_{12}, n_{21}, n_{22}) \sim Mult(N, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$

b) The logarithm of the counts its assumed to be additive and composed of an overall mean μ, and a mean value for each [redacted] are of the levels of the two variables.

c) They have a significant chance of having heart disease

d) $odds = e^{\beta_0 - \beta_1} = 1.91$

Question 6

a) Yes. Transform UEMP, MAN, LIC

b) Good model for prediction (high R-squared)
Low significance for regression coefficients ⟹ do variable selection
Some variables are highly correlated

c) The data set is very small. Possible to use backward selection or LOOCV

d) Much more parsimonious model: less variables, more significant and still high predictive power.

Question 7

a) For explanation read CART notes. The method generats rectangular areas defined by different latitudes and longitudes where the mean price value of houses is very similar

b) Linear regression won't generate areas but will assign predicted prices for each house

c) Cluster analysis