STAT 231 Tutorial Test 3

Monday March 24 4:30-5:20 p.m.

Version 1 Solutions

1. [8] Suppose $y_1, y_2, \ldots, y_n$ is an observed random sample from a $Poisson\,(\theta)$ distribution.

(a) Show clearly that the likelihood ratio test statistic for testing $H_0 : \theta = \theta_0$ is given by:

$$\Lambda(\theta_0) = 2n\left[\tilde{\theta}\log\left(\frac{\tilde{\theta}}{\theta_0}\right) + \theta_0 - \tilde{\theta}\right] \quad \text{where } \tilde{\theta} = \bar{Y}.$$

$$L(\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} = \frac{\theta^{n\bar{y}} e^{-n\theta}}{\prod_{i=1}^{n} y_i!} \quad \text{or more simply} \quad L(\theta) = \theta^{n\bar{y}} e^{-n\theta}, \quad \theta > 0$$

$$l(\theta) = \log L(\theta) = n\bar{y}\log\theta - n\theta, \quad \theta > 0$$

Thus

$$\begin{aligned}
\Lambda(\theta_0) &= 2\left[l\left(\tilde{\theta}\right) - l(\theta_0)\right] \\
&= 2\left[n\bar{Y}\log\tilde{\theta} - n\tilde{\theta} - n\bar{Y}\log\theta_0 + n\theta_0\right] \\
&= 2\left[n\tilde{\theta}\log\left(\frac{\tilde{\theta}}{\theta_0}\right) + n\left(\theta_0 - \tilde{\theta}\right)\right] \quad \text{since } \tilde{\theta} = \bar{Y} \\
&= 2n\left[\tilde{\theta}\log\left(\frac{\tilde{\theta}}{\theta_0}\right) + \theta_0 - \tilde{\theta}\right] \quad \text{as required.}
\end{aligned}$$

(b) Use this test statistic to test $H_0 : \theta = 2$ if $n = 10$ and $\sum_{i=1}^{10} y_i = 18$. **Write your final answers in the space provided.**

(i) The observed value of the test statistic is 0.2070.

$$\begin{aligned}
\lambda(\theta_0) &= 2n\left[\hat{\theta}\log\left(\frac{\hat{\theta}}{\theta_0}\right) + \theta_0 - \hat{\theta}\right] \quad \text{where } n = 10,\ \hat{\theta} = \frac{18}{10} = 1.8 \text{ and } \theta_0 = 2 \\
&= 2(10)\left[1.8\log\left(\frac{1.8}{2}\right) + 2 - 1.8\right] \\
&= 0.2070
\end{aligned}$$

(ii) The approximate $p-value$ using Normal tables is 0.65272.

$$\begin{aligned}
p-value &\approx P(W \geq 0.2070) \quad \text{where } W \backsim \chi^2(1) \\
&= 2\left[1 - P\left(Z \leq \sqrt{0.2070}\right)\right] \quad \text{where } Z \backsim G(0,1) \\
&= 2\left[1 - P(Z \leq 0.45)\right] = 0.65272
\end{aligned}$$

(iii) Your conclusion regarding $H_0 : \theta = 2$ is:

There is no evidence based on the data to contradict $H_0 : \theta = 2$.

2. [13] The following data were obtained from an experiment involving a chemical process in which the yield (y) in grams of the process was thought to be related to the reaction temperature (x) in degrees celsius:

i	x_i	y_i	i	x_i	y_i	i	x_i	y_i
1	50	108	11	72	160	21	93	204
2	53	118	12	74	161	22	94	208
3	54	130	13	75	161	23	95	204
4	55	124	14	76	168	24	97	211
5	56	130	15	79	174	25	100	218
6	59	141	16	80	175			
7	62	137	17	82	180			
8	65	143	18	85	183			
9	67	149	19	87	193			
10	71	161	20	90	188			

$$\sum_{i=1}^{25} x_i = 1871$$

$$\sum_{i=1}^{25} y_i = 4129$$

$$S_{xx} = 5679.36$$

$$S_{xy} = 11501.64$$

$$S_{yy} = 23629.36$$

To analyse these data (x_i, y_i), $i = 1, 2, \ldots, 25$ the simple linear regression model

$$Y_i \backsim G(\alpha + \beta x_i, \sigma) \qquad i = 1, 2, \ldots, 25 \text{ independently}$$

is assumed where α, β and σ are unknown parameters and the x_i's are known constants.

Summary of Distributions for Simple Linear Regression

Random variable	**Distribution**	**Mean or df**	**Standard Deviation**
$\tilde{\beta} = \frac{S_{xy}}{S_{xx}}$	Gaussian	β	$\sigma \left[\frac{1}{S_{xx}}\right]^{1/2}$
$\frac{\tilde{\beta}-\beta}{S_e/\sqrt{S_{xx}}}$ $S_e^2 = \frac{1}{n-2}\left(S_{yy} - \tilde{\beta} S_{xy}\right)$	t	$\text{df} = n-2$	
$\tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{x}$	Gaussian	α	$\sigma \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]^{1/2}$
$\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta} x$	Gaussian	$\mu(x) = \alpha + \beta x$	$\sigma \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]^{1/2}$
$\frac{\tilde{\mu}(x)-\mu(x)}{S_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}}$	t	$\text{df} = n-2$	
$Y - \tilde{\mu}(x)$	Gaussian	0	$\sigma \left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]^{1/2}$
$\frac{Y-\tilde{\mu}(x)}{S_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}}$	t	$\text{df} = n-2$	
$\frac{(n-2)S_e^2}{\sigma^2}$	Chi-squared	$\text{df} = n-2$	

Write your final answer only in the space provided.

(a) The least squares estimate of β is $\underline{\frac{11501.64}{5679.36} = 2.0252}$.

(b) The least squares estimate of α is $\underline{\frac{4129}{25} - \left(\frac{11501.64}{5679.36}\right)\left(\frac{1871}{25}\right) = 13.5967}$.

(c) An unbiased estimate of σ^2 is $\underline{s_e^2 = \frac{1}{23}\left[23629.36 - \left(\frac{11501.64}{5679.36}\right)(23629.36)\right] = 14.6367 = (3.8258)^2}$.

(d) The $p-value$ for testing the hypothesis of no relationship between yield and temperature $(H_0 : \beta = 0)$

is approximately equal to $\underline{\quad 0 \quad}$.

$$d = \frac{\left|\hat{\beta} - 0\right|}{s_e/\sqrt{S_{xx}}} = \frac{|2.0252 - 0|}{3.8258/\sqrt{5679.36}} = 39.89$$

$$p - value = 2\left[1 - P(T \leq 39.89)\right] \approx 0 \quad \text{where} \quad T \backsim t\,(23)$$

(e) A 95% confidence interval for the mean response at a temperature of $x = 60$ is $\underline{[132.89, 137.33]}$.

$$(13.5967) + (2.0252)\,(60) \pm (2.0687)\,(3.8258)\sqrt{\frac{1}{25} + \frac{(60 - 74.84)^2}{5679.36}}$$
$$= 135.1066 \pm 2.2213$$
$$= [132.8852, 137.3279]$$

(f) A 95% prediction interval for the response at a temperature of $x = 40$ is $\underline{[85.74, 103.46]}$.

$$(13.5967) + (2.0252)\,(40) \pm (2.0687)\,(3.8258)\sqrt{1 + \frac{1}{25} + \frac{(40 - 74.84)^2}{5679.36}}$$
$$= 94.6033 \pm 8.8616$$
$$= [85.7417, 103.4648]$$

(g) What warning would you give regarding the interval in (f)?

The value $x = 40$ is outside the observed interval of x values $[50, 100]$. Therefore the prediction interval is based on an assumption that the linear relationship holds below $x = 50$ and we have no data to support this assumption.

3. [4] To analyse the data (x_i, y_i), $i = 1, 2, \ldots, 100$ the simple linear regression model

$$Y_i \backsim G(\alpha + \beta x_i, \sigma) \qquad i = 1, 2, \ldots, 100 \text{ independently}$$

is assumed where α, β and σ are unknown parameters and the x_i's are known constants.

Use all **three** of the following plots to make a conclusion regarding the reasonableness of the model assumptions. If the assumed model is not reasonable suggest a better model.

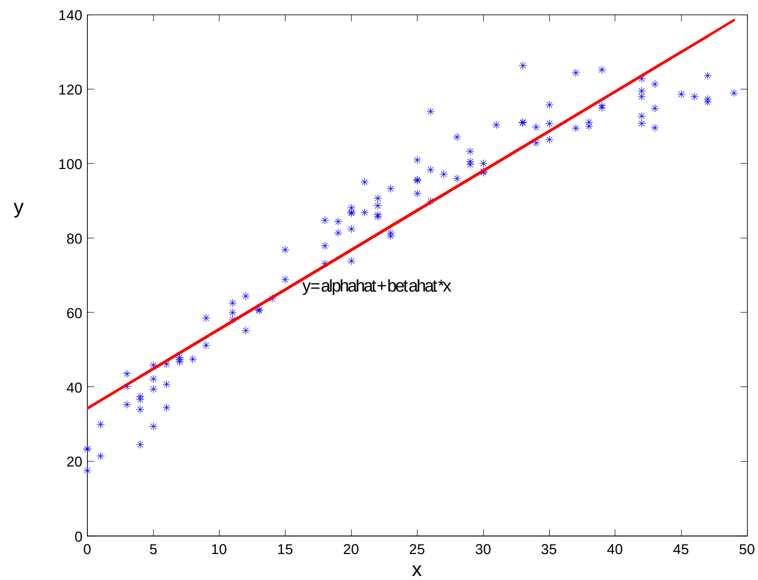
If the Gaussian linear model with constant variance is reasonable then we expect to see that a scatterplot of the data lie reasonably along the fitted line. Plot A indicates that the linear model is not adequate since for small and large values of x the points lie below the fitted line and for values of x in the middle the points all lie above the fitted line.

If the Gaussian linear model with constant variance is reasonable then we expect to see the points in a standardized residual plot lying in a horizontal band around the line $r_i^* = 0$. The residual plot in Plot B does not exhibit this behaviour since for small and large values of x the residuals are all negative and for values of x in the middle the residuals are all positive. Plot B indicates the assumptions do not hold.

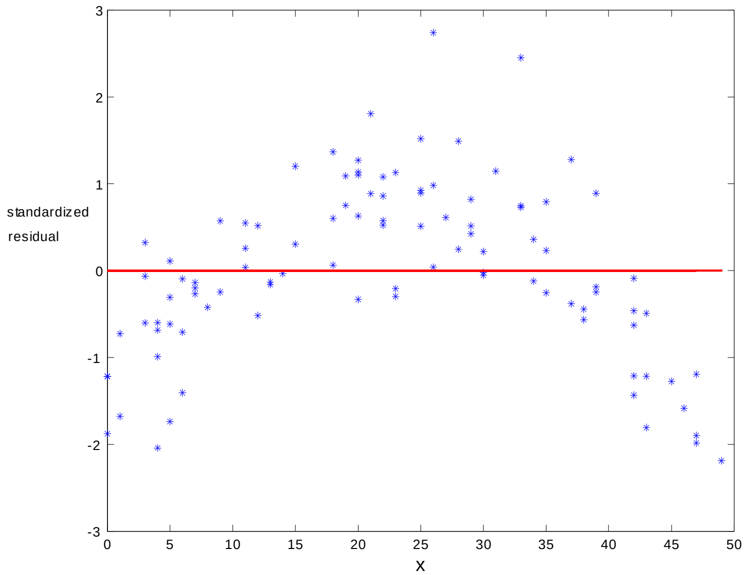
If the Gaussian linear model with constant variance is reasonable then we expect the points in a qqplot of the residuals to lie along a straight. In Plot C we see that, although the points in the middle lie along a straight line, for small and large values of the standard Normal quantiles the points all lie below the line. Plot C also indicates the assumptions do not hold.

The pattern of departures observed in Plots A and B suggest that a quadratic model for the mean of the form $\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ would provide a better fit to the data.

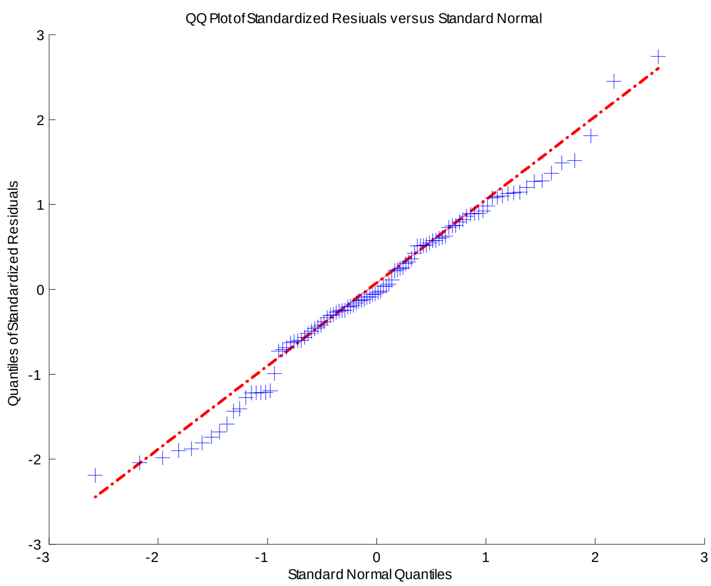
Plot A:



Plot B:


standardized
residual
X
3
2
1
0
-1
-2
-3
0
5
10
15
20
25
30
35
40
45
50

Plot C:


QQ Plot of Standardized Resiuals versus Standard Normal
Quantiles of Standardized Residuals
Standard Normal Quantiles
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3