Theory Exam 1

36-401

8 October 2015

DO NOT START UNTIL 3 PM

Record all of your answers in the blue-book provided; if you need more space, ask for another blue-book. Show work for all problems; even a completely correct answer will receive no credit if unsupported by work.

No electronic devices of any kind are needed for this exam, or permitted. Tables at the end of the exam give all necessary values for special functions.

You are allowed a formula sheet of one side one 8.5×11 inch piece of paper.

1. *Why $n-2$?* In this problem, we will show why an unbiased estimator of σ^2 is not the MLE, $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$, but rather $\frac{n}{n-2}\hat{\sigma}^2$.

   Throughout this problem, you may presume that all the assumptions of the simple linear regression model hold. If at any point you need to assume the noise variables ϵ_i are Gaussian and independent, explain why.

   Notation $\bar{x}, \bar{y}$: sample means of x and y; s_X^2, sample variance of x, $n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

   (a) (15) Show that the estimates of the slope and intercept can be written as the true coefficients plus a weighted sum of noise terms:

$$\begin{aligned}\hat{\beta}_0 &= \beta_0 + \sum_{j=1}^{n}\left(\frac{1}{n} - \overline{x}\frac{x_j - \overline{x}}{ns_X^2}\right)\epsilon_j \\ \hat{\beta}_1 &= \beta_1 + \sum_{j=1}^{n}\frac{x_j - \overline{x}}{ns_X^2}\epsilon_j\end{aligned}$$

   Hints: Start from formulas for the estimated coefficients in terms of the data, and use one of the modeling assumptions; get $\hat{\beta}_1$ first.

   (b) (5) Show that the estimated conditional expectation at an arbitrary x can be written as the true expected value plus a weighted sum of noise terms:

$$\hat{\beta}_0 + \hat{\beta}_1 x = \beta_0 + \beta_1 x + \sum_{j=1}^{n}\left(\frac{1}{n} + (x - \bar{x})\frac{x_j - \bar{x}}{ns_X^2}\right)\epsilon_j$$

   (c) (5) The "Kronecker delta" symbol, δ_{ij} is 1 when $i = j$ and 0 otherwise. Show that the i^{th} residual (i.e., the residual at $x = x_i$) can be written as a weighted sum of the noise terms:

$$e_i = \sum_j \left(\delta_{ij} - \frac{1}{n} - (x_i - \bar{x})\frac{x_j - \bar{x}}{ns_X^2}\right)\epsilon_j$$

   (d) (10) Show that

$$\mathbb{E}\left[e_i^2\right] = \sigma^2\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{ns_X^2}\right)$$

   Describe, *qualitatively*, how $\mathbb{E}\left[e_i^2\right]$ varies with x_i.

   (e) (5) Show that $\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{n-2}{n}\sigma^2$.

2. *Confidence interval for σ^2* As you remember, the sum of squared errors is $SSE = \sum_{i=1}^{n} e_i^2$. We know that in the Gausian-noise simple linear regression model, the ratio $SSE/\sigma^2 \sim \chi^2_{n-2}$.

   (a) (5) Given σ^2 and a number $\alpha > 0$, find a formula for an interval which contains SSE with probability $1 - \alpha$, i.e., numbers l and u such that

   $$\Pr(l \leq SSE \leq u) = 1 - \alpha$$

   Express your answer in terms of n, σ^2, and the quantiles of χ^2 distributions.

   (b) (5) Using your answer from 2a, find a formula for a $1 - \alpha$ confidence interval for σ^2. Your upper and lower limits should be expressed in terms of SSE, n, and the quantiles of χ^2 distributions.

   (c) (5) We run a simple linear regression model with 43 observations and obtain a sum of squared errors of 100. Find a 95% confidence interval for σ^2.

   (d) (5) With the data from 2c, can we reject the null hypothesis that $\sigma^2 = 3$? Explain.

3. *Interpreting regression output* The following regression output was obtained using the city-economy data set used in homework 3. Recall that for each of 366 cities in the US, this records the city's per-capita gross metropolitan product, in dollars per person per year, and its population.

```
##
## Call:
## lm(formula = pcgmp ~ log10(pop), data = bea)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21572  -4765  -1016   3686  40207
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -23306       4957    -4.7  3.7e-06
## log10(pop)     10246        900    11.4  < 2e-16
##
## Residual standard error: 7930 on 364 degrees of freedom
## Multiple R-squared:  0.263,Adjusted R-squared:  0.26
## F-statistic:  130 on 1 and 364 DF,  p-value: <2e-16
```

For the following questions, please explain clearly which parts of the output are the basis for your answers. Include the units of variables wherever possible. *General hint:* Use the tables at the end of this exam.

(a) (5) What is the predictor variable? What is the response variable? Which variables were transformed, and how?

(b) (5) Write the equation for the estimated conditional mean function; use numerical values rather than symbols like $\hat{\beta}_0$.

(c) (5) According to the estimated model, what is the average per-capita gross metropolitan product of cities with a population of one million people? Of cities with a population of two hundred thousand people? Do these numbers seem reasonable?

(d) (5) Based on the estimated coefficients, can you give an estimate of $\mathbb{E}[Y|X = 0]$? If yes, what is it (and show your work); if not, explain why not (missing information, inappropriate assumptions, etc.).

(e) (5) Give a 99% confidence interval for β_1, assuming all the model assumptions hold.

(f) (5) What is $\hat{\sigma}^2$, the in-sample mean-squared error?

(g) (5) From n, the standard error of $\hat{\beta}_1$, and $\hat{\sigma}^2$, can you find the sample variance in population across cities? If so, what is it? If not, explain.

(h) (5) Which part (or parts) of the output (if any) tests the assumption that the relationship between the (possibly transformed) predictor variable and the (possibly transformed) response variable is linear?

	20	21	40	41	42	43	44	45	84	86	364	366
0.005	7.43	8.03	20.7	21.4	22.1	22.9	23.6	24.3	54.4	56.0	298	300
0.025	9.59	10.30	24.4	25.2	26.0	26.8	27.6	28.4	60.5	62.2	313	315
0.05	10.90	11.60	26.5	27.3	28.1	29.0	29.8	30.6	63.9	65.6	321	323
0.1	12.40	13.20	29.1	29.9	30.8	31.6	32.5	33.4	67.9	69.7	330	332
0.9	28.40	29.60	51.8	52.9	54.1	55.2	56.4	57.5	101.0	103.0	399	401
0.95	31.40	32.70	55.8	56.9	58.1	59.3	60.5	61.7	106.0	109.0	409	412
0.975	34.20	35.50	59.3	60.6	61.8	63.0	64.2	65.4	111.0	114.0	419	421
0.995	40.00	41.40	66.8	68.1	69.3	70.6	71.9	73.2	121.0	124.0	437	439

Table 1: Selected quantiles of χ^2 distributions: the probabilities are given by the rows, and the number of degrees of freedom by the columns.

	20	21	40	41	42	43	44	45	84	86	364	366	Inf
0.005	-2.85	-2.83	-2.70	-2.70	-2.70	-2.70	-2.69	-2.69	-2.64	-2.63	-2.59	-2.59	-2.58
0.025	-2.09	-2.08	-2.02	-2.02	-2.02	-2.02	-2.02	-2.01	-1.99	-1.99	-1.97	-1.97	-1.96
0.05	-1.72	-1.72	-1.68	-1.68	-1.68	-1.68	-1.68	-1.68	-1.66	-1.66	-1.65	-1.65	-1.64
0.1	-1.33	-1.32	-1.30	-1.30	-1.30	-1.30	-1.30	-1.30	-1.29	-1.29	-1.28	-1.28	-1.28
0.9	1.33	1.32	1.30	1.30	1.30	1.30	1.30	1.30	1.29	1.29	1.28	1.28	1.28
0.95	1.72	1.72	1.68	1.68	1.68	1.68	1.68	1.68	1.66	1.66	1.65	1.65	1.64
0.975	2.09	2.08	2.02	2.02	2.02	2.02	2.02	2.01	1.99	1.99	1.97	1.97	1.96
0.995	2.85	2.83	2.70	2.70	2.70	2.70	2.69	2.69	2.64	2.63	2.59	2.59	2.58

Table 2: Selected quantiles of t distributions, with selected degrees of freedom; the last column gives quantiles of the z distribution.

x	log(x)	log10(x)	(x^(0.1)-1)/(0.1)
1	0.000	0.000	0.000
2	0.693	0.301	0.718
3	1.100	0.477	1.160
4	1.390	0.602	1.490
5	1.610	0.699	1.750
6	1.790	0.778	1.960
7	1.950	0.845	2.150
8	2.080	0.903	2.310
9	2.200	0.954	2.460
10	2.300	1.000	2.590
1000	6.910	3.000	9.950

Table 3: Selected values of selected transformations.