36-401 Test 1

There are three questions. 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.

Write in pen.

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

The grading scheme is:
Question 1: 40
Question 2: 20
Question 3: 40

(1) Suppose we have data

$$(X_1, Y_1), \ldots, (X_n, Y_n).$$

Let

$$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i, \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i, \quad s_X^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2.$$

Throughout this question, you can assume that the linear model is correct. You can also assume that the noise variables ϵ_i are Gaussian.

(a) (15) Show that the least squares estimators can be written as

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

(b) (5) Let $m(x) = \beta_0 + \beta_1 x$ and $\widehat{m}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$. Show that

$$\widehat{m}(x) = m(x) + \sum_{j=1}^n \left(\frac{1}{n} + (x - \overline{X})\frac{X_j - \overline{X}}{ns_X^2}\right)\epsilon_j.$$

(c) (5) Define δ_{ij} to be 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

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

(d) (10) Show that

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

(e) (5) Show that $\mathbb{E}\left[\widehat{\sigma}^2\right] = \frac{n-2}{n}\sigma^2$ where $\widehat{\sigma}^2 = (1/n)\sum_i e_i^2$.

(2) Define the sum of squared errors by $SSE = \sum_{i=1}^{n} e_i^2$ where $e_i = Y_i - [\widehat{\beta}_0 + \widehat{\beta}_1 X_i]$. Recall that, in the Gaussian-noise simple linear regression model, we have that $SSE/\sigma^2 \sim \chi^2_{n-2}$.

(a) (10) Find a $1 - \alpha$ confidence interval for σ^2. Express your answer in terms of SSE, n and χ^2_{n-2} quantiles.

(b) (5) Suppose that 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.

(c) (5) With the data from part (b), can we reject the null hypothesis that $\sigma^2 = 3$ at level $\alpha = 0.05$? Explain.

(3) The following regression output was obtained using the city-economy data set. 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.

```
x = log10(pop)   ### log10 computes log to the base 10. For example, log10(100) = 2.
y = pcgmp
out = lm(y ~ x)
summary(out)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## 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
## x              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, explain clearly which parts of the output are the basis for your answers. If necessary, use the tables at the end of this exam.

(a) (5) What is the predictor (covariate) 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 $\widehat{\beta}_0$.

(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.

(e) (5) Give a 95% confidence interval for β_1, assuming all the model assumptions hold.

(f) (5) What is $\widehat{\sigma}^2$? (You may use either of the estimators that we discussed.)

(g) (5) Can you find the sample variance of the variable pop from the information in the output? 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 predictor variable and the 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 left-tail quantiles of χ^2 distributions: the probabilities are given by the rows, and the number of degrees of freedom by the columns. For example $P(\chi^2_{20} < 7.43) = 0.005$ and $P(\chi^2_{21} < 41.40) = 0.995$.

	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. For example, $P(T < 2.85) = 0.995$ if $T \sim t_{20}$.

x	log(x)	log10(x)	$(x^{1/10} - 1)/(0.1)$
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

Table 3: Selected values of some transformations.