Chicago Booth BUS 41100
Solutions to Final Exam **SAMPLE #2**

Instructor: Max H. Farrell

These solutions are a guide only! Your answers should show more work/detail/reasoning.

1 Short Answer & Multiple Choice

(a) The slope is clearly negative, eliminating (i) and (v). The intercept is certainly not 12, eliminating (iii). Examining the axis shows the slope is -2.5, giving (ii).

(b) The first three are true, this comes from the lecture slides. Choice (iv) is false: Y can't have the same correlation with all the X variables individually, not does it make sense to somehow be correlated with all of them.

(c) Choice (i) is only true for linear regression. Choices (ii) and (iii) are false by definition of an odds ratio. Only Choice (iv) is correct.

(d) There is a clear periodic pattern and a clear upward time trend, thus (ii) and (iii) are needed. There is not obvious pattern for lagged values, e.g. repeated above-trend peaks or repeated deviations from the cyclical pattern, so lagged outcomes are not going to add anything but noise.

(e) Choice (iii) is the only correct one. From the lecture slides: the log odds goes up by 2.3 for a one unit change in `x1`, which means the odds ratio itself is multiplied by $e^{b_1} = e^{2.3} = 10$ for a one unit change in `x1`. If this one is correct, the others must be false by definition. That is, if you work through what an odds ratio is, you'll see that none of the other changes are possible.

2 Multiple Linear Regression

(a) Tropicana positions itself as the premium brand, while the store brand is the budget option. Minute Maid attempts to slot in between. We see from the correlations that the prices mostly move together, which is expected (e.g. as the raw ingredient of oranges changes price) and from the plots we can see that the market positions are quite strictly kept to, that is, Minute Maid is hardly ever cheaper than the store brand, and hardly ever more dear than Tropicana.

(b) This is a log-log regression of sales on price, and hence the slope coefficient is the price elasticity. The numerical value of -2.64 means that when the price goes up by 1%, sales fall by 2.64%.

(c) This is a log-log regression of sales on prices, and hence the slope coefficients are all price elasticities. The two for `log(tropicana.price)` and `log(store.brand.price)` are the elasticities with respect to competitors prices (a cross price elasticity). The values are positive, meaning that when competitors prices go up, Minute Maid sales go up, by 1.6% and 1.29% respectively, for a 1% rise in price.

(d) **(i)** This value represents a shift in the intercept of the demand curve, so all else held equal, having a sale means a 189% increase in sales. At least the direction makes sense: having an ad means higher sales. The magnitude seems huge though.

(ii) This value represents a change in the slope for when an ad is present, i.e. a change in the price elasticity. Consumers become more elastic. One possible reason for this is that if you have an ad you raise awareness of your product and the price reduction, and are more likely to buy. The estimate value means that having a 1% fall in price gives a 3.14% rise in sales with no ad, but coupled with an ad gives a 4% increase in sales.

3 Model Building

(a) **(i)** Each row shows the partial F test of one model against the next most complicated model. From each row we see that, according to this metric, each new model represents a worthwhile addition. That is, Model 6 is the best, according to this.

(ii) The problem is that it relies on hypothesis testing and R^2, i.e. model fit. From the latter, it is prone to overfitting, because R^2 is not a good goal in and of itself, prediction is the goal. From the former, it is prone to the multiple testing problem, and the choosing a specific direction.

(b) **(i)** This is stepwise selection based on BIC. The process begins from the null model, which is empty. At each step one variable is added, and the variable that leads to the greatest improvement in BIC is kept. This is repeated until no variables lead to improvement. Main effects are tested first, then interactions.

(ii) There are **exactly** four: 1) the starting place (here the empty model), 2) the scope of the search (the full model), 3) the direction of the search (forward), and 4) the criteria to evaluate each step (BIC). Nothing else is controlled by the user.

(c) **(i)** The difference is the starting point only, we force the stepwise process to begin with 3 variables already in it, and the rest of the search proceeds as above.

(ii) The reason is that age and experience are highly correlated, in fact nearly perfectly: the correlation is 0.98. So they are very close to being the same exact variable, and so close that BIC can't tell them apart. Even without the numerical correlation, it is conceptually clear that they would be correlated: no 30 year old has 20 years of experience. Since the only difference between the final selected models is that one includes age and the other includes experience, there is no other possible answer. In particular, note that age was added in the third step in part (c), so it is the most important variable (BIC-wise) after sex and education, meaning that it would be added in step 1 in part (c) if it were different enough from experience. Note that it is wrong to suggested that experience/age don't matter once the other variables are conditioned upon: if this were true it would not be selected first. In particular, it's not true that once adding experience squared, the main effect doesn't matter, because in part (b) age was selected before experience squared.

(iii) It could be either, or no preference. We prefer part (e) for the reason that including experience squared without the main effect is a bad idea. Even though experience and age are nearly perfectly correlated, and so the models are really the same, interpretation of the model in part (d) is more difficult on the face of it. Overall, part (e)'s model is more conceptually well-grounded. However, given that the models are the same, it really doesn't matter as long as you are careful with the interpretation. One reason to prefer the model in (d) is that it is entirely based on BIC, and so if we really had no idea what to use to predict wages, this might be a better way to go.

(d) **(i)** The final two lines compute the AIC and BIC, respectively, for the two different models. To interpret the output: the lower the number the better is the model, according to that specific measure.

(ii) The penultimate line shows that the AIC for the `job.type` regression is slightly lower, so that is preferred over the collapsed `job.level` measure. The last line shows the reverse conclusion: the BIC is lower for the `job.level` regression. The reason for this is that, as discussed in class, AIC tends to prefer larger models and BIC likes smaller models. So there is some additional information in using `job.type` rather than the coarser `job.level`, and AIC picks that up, whereas for BIC the additional information is not worth the extra complexity.

4 Time Series Regression

This problem studies the time series of returns for the Istanbul Stock Exchange. We have daily returns for an index variable for this exchange (`ISE`) from January 5th, 2009 to February 22, 2011, along with returns for indexes for the following markets: US (`SP`), Germany (`DAX`), the UK (`FTSE`), Japan (`NIKKEI`), Brazil (`BOVESPA`), and the MSCI Emerging markets index (`EM`).

(a) **(i)** The ACF plot shows the estimates of the autocorrelation function: the first point is the sample correlation of the series with the current value, 1 by definition, the second is the correlation with the first lag, and so on. The plot shows that there is almost no time-series dependence in this series.

(ii) The model is $Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t$, with $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.

(iii) The fact that the first lag is not significantly related to the current value matches the fact that the ACF plot shows no time dependence. The coefficient estimate is positive, as shown in the plot.

(b) **(i)** The BIC values are essentially ordered the same as the pairwise correlations. The best (lowest) BIC value matches the highest in-sample correlation.

(ii) We can't tell. All we have is the correlations, we don't have the sample standard deviations of any of the series, nor the residual variance estimate. It's not even true that the regression using `EM` is the "most likely" to be significant, because we don't know how the spread in this X value compares to the others, so even though it has the highest correlation, a different value could be significant. This is all another way of saying we can't judge anything based on R^2 alone.

5 Transformations

(a) $\mathbb{E}[\texttt{brain.weight} \mid \texttt{body.weight}] = e^{\beta_0}(\texttt{body.weight})^{\beta}_{1}$

(b) One justification that requires no additional evidence is that we might want to know about the elasticity. We have a lot of different mammals, and looking at % changes levels the playing field. A 1 kg increase in body weight means a lot more to a guinea pig than an African elephant. This is exactly like the GDP and imports example from class.

Alternatively we could look at a level-level model and study the spread of X values to justify taking $\log(X)$ or nonconstant variance for $\log(Y)$. Just like in the week 3 & 4 slides.

(c) 92% of the variance is explained (see the R^2). We do have reason to believe a linear relationship between this variables: the t-statistic is bigger than two, so we reject the null $H_0 : \beta_1 = 0$ at the 5% level in favor of the two sided alternative.

(d) The fitted value is $e^{5.67}$. The interval is roughly $[e^{4.25}, e^{7.1}]$.

6 Logistic Regression

For 546 homes we observe the following informaiton

`AC` = 1 if it has air conditioning, 0 if not,
`price` = sale price, and
`bedrooms` = number of bedrooms.

Our goal is to predict which houses have air condition and which do not.

(a) First we use a logistic regression with only `bedrooms` and obtain the following output.

```
Call:
glm(formula = AC ~ bedrooms, family = "binomial")

Coefficients:
            Estimate Std. Error z value    Pr(>|z|)
(Intercept)   -2.179      0.399   -5.46 0.000000047
bedrooms       0.469      0.127    3.68     0.00023
```

Give a precise, numerical interpretation of the relationship between `bedrooms` and `AC` by:

(i) Interpreting the coefficient estimate, 0.469, in this context.

Solution. *See Week 6.*

(ii) Interpreting the statistical significance as best you can with this output.

Solution. *The most useful thing to do is probably the usual thing: test the null of* $\beta_1 = 0$*, which we reject because* $0.469/0.127 > 2$*. So we conclude that at the 5% level bedrooms are associated with an increase in the probability of having AC.*

(b) We now switch to using `price` and obtain the following output.

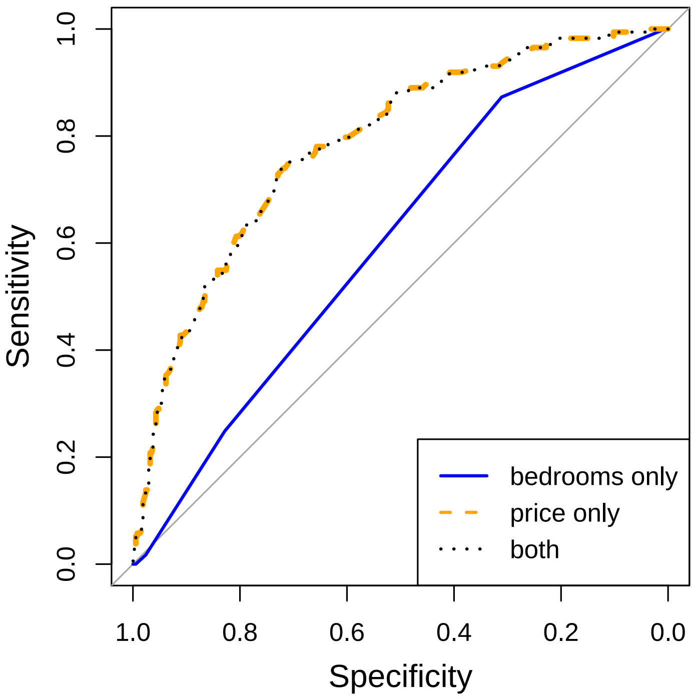
```
Call:
glm(formula = AC ~ price, family = "binomial")

Coefficients:
               Estimate  Std. Error z value            Pr(>|z|)
(Intercept) -3.76328283  0.34599029   -10.9 <0.0000000000000002
price        0.00004223  0.00000459     9.2 <0.0000000000000002
```

Based on this output, is `price` or `bedrooms` better at predicting whether or not a house has AC? Cite specific numeric evidence in favor or your argument.

Solution. *There is really no way to tell from this output. Both are statistically significant.*

Use the plot below to answer the next two questions.



(c) Provide an intuitive explanation of what is being plotted here.

Solution. *See Week 6.*

(d) Based on this plot, comment on how `bedrooms` contribute to predicting `AC` when used in conjunction with `price`.

Solution. *We can see that* `bedrooms` *adds nothing to the model's ability to classify houses into AC yes/no.*