

## Q2

24 Points

Determine whether the following statements are TRUE or FALSE.

### Q2.1

2 Points

It is possible for a Kaplan-Meier estimate of survival to be exactly zero.

### Q2.2

2 Points

Adding additional predictors to a linear regression model will never decrease the  $R^2$  for that model.

### Q2.3

2 Points

A linear model with many highly significant predictors guarantees a high  $R^2$  value.

### Q2.4

2 Points

If the outcome variable in a linear regression model is not normally distributed, then we automatically know that the assumptions are not satisfied.

### Q2.5

2 Points

If the assumptions for a linear regression model are satisfied, then the difference between the observed and predicted outcomes must be normally distributed.

### Q2.6

2 Points

For a logistic regression model with a single continuous predictor, a positive slope parameter implies that higher values of this predictor are associated with a higher probability of the outcome occurring.

**Q2.7**

2 Points

If we know the probability of an event's occurrence, then we automatically know the odds of that event occurring as well.

**Q2.8**

2 Points

The logistic regression model is a linear model relating predictors to the odds of an event occurring.

**NOTE: WE DIDN'T COVER Q2.9 THIS SEMESTER.**

**Q2.9**

2 Points

In general, k-NN models will have higher classification accuracy if all predictors are on the same scale.

**Q2.10**

2 Points

A negative slope parameter in a multiple regression model implies that lower values of the predictor corresponding to that parameter are associated with higher values of the outcome.

**Q2.11**

2 Points

If the overall F statistic for a multiple regression model is significant, then we know that at least one of the t-tests for the predictors in the model will be statistically significant.

**Q2.12**

2 Points

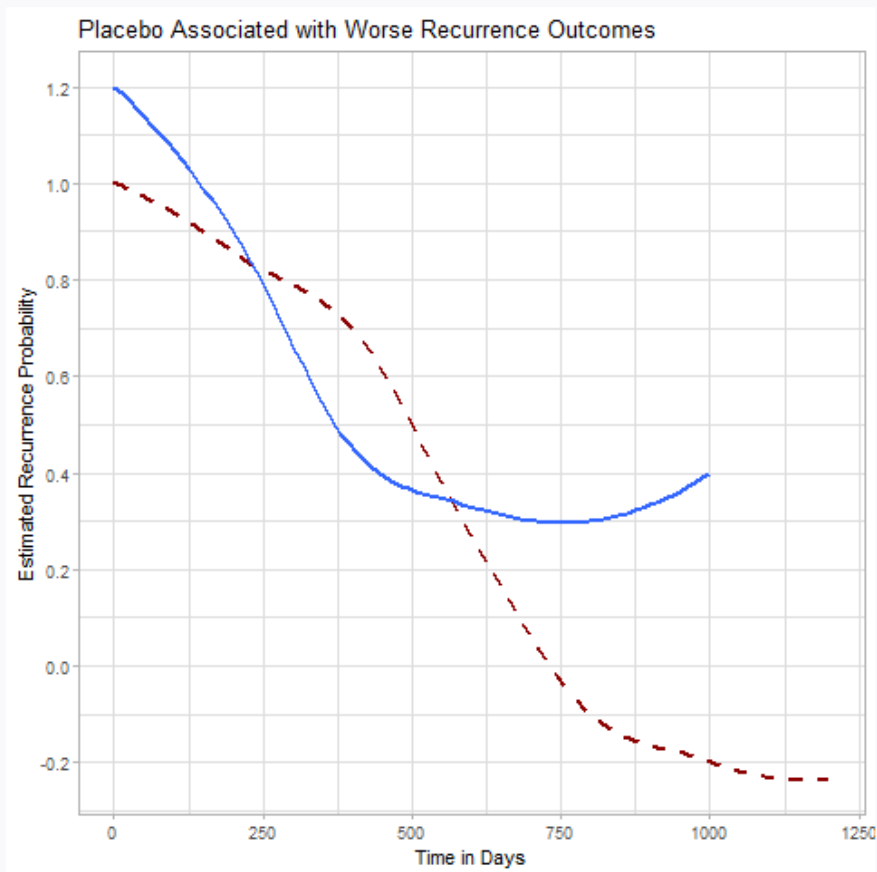
The predictors in a linear regression model must be normally distributed in order for our assumptions to hold.

### Q3

12 Points

Researchers at a pharmaceutical firm developed a chemotherapy agent for a certain rare cancer and conducted a placebo-controlled clinical trial to investigate the efficacy of the drug. The primary outcome was the time to recurrence of the cancer.

The investigators created a Kaplan-Meier curve comparing these two drugs. In their plot, placebo is given by a solid blue line and the chemotherapy agent in the dashed red line. However, as the data scientist on the team, you suspect that the Kaplan-Meier plot was created incorrectly (provided below).



### Q3.1

8 Points

Comprehensively identify everything wrong with the Kaplan-Meier plot they created.

### Q3.2

4 Points

Assume the Kaplan-Meier curves are accurate (they emphatically are not). What are the median survival times in the placebo and chemotherapy groups?

## Q4

26 Points

Vinho verde is a Portuguese wine style known for its light and refreshing character. Wine producers are interested in predicting the quality of their wine based on its chemical properties. The `wine` dataset contains 1,000 observations of white vinho verde style wines, some chemical properties of the wine, as well as a numeric quality score based on its rating by professional oenologists (wine specialists). These data were modified from Cortez et al. (Decis. Support Syst., 2009).

- `citric_acid`: Citric acid content in  $\text{g/dm}^3$
- `sugar`: Residual sugar in  $\text{g/m}^3$
- `chlorides`: Sodium chloride (salt) in  $\text{g/dm}^3$
- `so2`: Total sulfur dioxide in  $\text{mg/dm}^3$
- `pH`: pH
- `alcohol`: Percent alcohol by volume
- `region`: Growing region (1, 2, or 3, each corresponding to a different growing region)
- `quality`: Quality score assigned by oenologists

#### Q4.1

4 Points

Fit a predictive model for quality score based on the predictors in the dataset. What advice might you give to vinho verde producers in order to maximize the perceived quality by oenologists? Explain, using results from your model.

If needed, you may upload any supporting work/code in the space provided.

 No files uploaded



#### Q4.2

8 Points

Comprehensively evaluate whether the assumptions for the linear model you created in Q4.1 are satisfied.

If needed, you may upload any supporting work/code in the space provided.

 No files uploaded




### Q4.3

4 Points

Interpret the intercept estimate and one slope estimate of your choice from your model in Q4.1.

If needed, you may upload any supporting work/code in the space provided.

 No files uploaded

### Q4.4

6 Points

Using results from your model in Q4.1, is there evidence that sugar is associated with quality score? Explain, using a formal hypothesis test.

If needed, you may upload any supporting work/code in the space provided.

 No files uploaded

### Q4.5

4 Points

Construct a 95% confidence interval corresponding to the intercept from your model in Q4.1. Is this confidence interval useful? Explain.

If needed, you may upload any supporting work/code in the space provided.

 No files uploaded

## Q5

38 Points

A major source of morbidity among liver transplant patients is post-transplant diabetes mellitus (PTDM). The data contained in `transplant.csv` are representative of retrospective patient-level data collected from a major US hospital system; each observation corresponds to a liver transplant patient. The following data were collected:


- `eti`: Reason (etiology) for liver transplantation: `cirr_alc` (alcoholic cirrhosis), `cirr_non` (non-alcoholic cirrhosis), or `other` (non-cirrhosis related)
- `sex`: Sex of the patient: M (male), or F (female)
- `bmi`: BMI of the patient in  $\text{kg/m}^2$
- `meld`: MELD score of the patient (a numeric score indicating severity of liver disease for transplant purposes; the higher, the more severe)
- `pre_dm`: Indicator for whether the patient had diabetes pre-transplantation
- `cold_isch`: Cold ischemia time in hours (cold ischemia time is the duration between when the liver is removed from the donor and transplanted into the recipient)
- `don_bmi`: BMI of the donor, in  $\text{kg/m}^2$
- `post_dm`: Indicator for development of PTDM within two years of the transplantation

### Q5.1

6 Points

Fit a logistic regression model for whether a patient develops PTDM within two years of transplant using all predictors in `transplant.csv`. Interpret the parameter estimates corresponding to `etiother`, `pre_dm`, and `cold_isch` *on the odds ratio scale*.

If needed, you may upload any supporting work/code in the space provided.

 No files uploaded


### Q5.2

8 Points

Consider a female patient with alcoholic cirrhosis who did not have diabetes prior to transplant. Suppose this patient's BMI and her donor's BMI were equal to 26.5 (the national average in the US), and the donor liver's cold ischemia time was equal to the average cold ischemia time in the dataset.

Using your model from Q5.1, what MELD scores must this patient have in order for her predicted probability of developing PTDM within two years of transplant to be less than 10%?

If needed, you may upload any supporting work/code in the space provided.

 No files uploaded




### Q5.3

5 Points

Adjusting for all other predictors, is there evidence that cold ischemia time is associated with developing PTDM within two years of transplant among patients in `transplant.csv`? Explain, using a formal hypothesis test.

If needed, you may upload any supporting work/code in the space provided.

 No files uploaded

### Q5.4

5 Points

Adjusting for all other predictors, is there evidence that the relationship between cold ischemia time and PTDM status depends on a patient's diabetes status prior to transplant? Explain, using a formal hypothesis test.

If needed, you may upload any supporting work/code in the space provided.

 No files uploaded

### Q5.5

5 Points


Researchers are interested in comparing classification accuracy of three different models.

Data from a new set of 50 patients at the same hospital system are available in

`new_patients.csv`

Using your model from Q5.1, predict whether the patients in `new_patients.csv` will develop PTDM within two years of their transplant. What is the prediction accuracy of this model?

If needed, you may upload any supporting work/code in the space provided.

 No files uploaded

**Note: we didn't cover 5.6/5.7.**


### Q5.6

5 Points

Consider the four continuous predictors only (BMI, MELD score, cold ischemia time, and donor BMI). Fit a k-NN model with  $k = 15$  using data from `transplant.csv`. Using this model, predict whether the patients in `new_patients.csv` will develop PTDM within two years of their transplant. What is the prediction accuracy of this model?

**Note: to save time, do not standardize the predictors.** (for the curious, prediction accuracy is 68% after the standardization).

If needed, you may upload any supporting work/code in the space provided.

 No files uploaded

### Q5.7

4 Points

Compare the results from Q5.5 and Q5.6. What do you think is the reason for any differences you observed in the prediction accuracy of the models?

If needed, you may upload any supporting work/code in the space provided.

 No files uploaded