Homework 01

Spencer Pease

1/25/2021

Questions

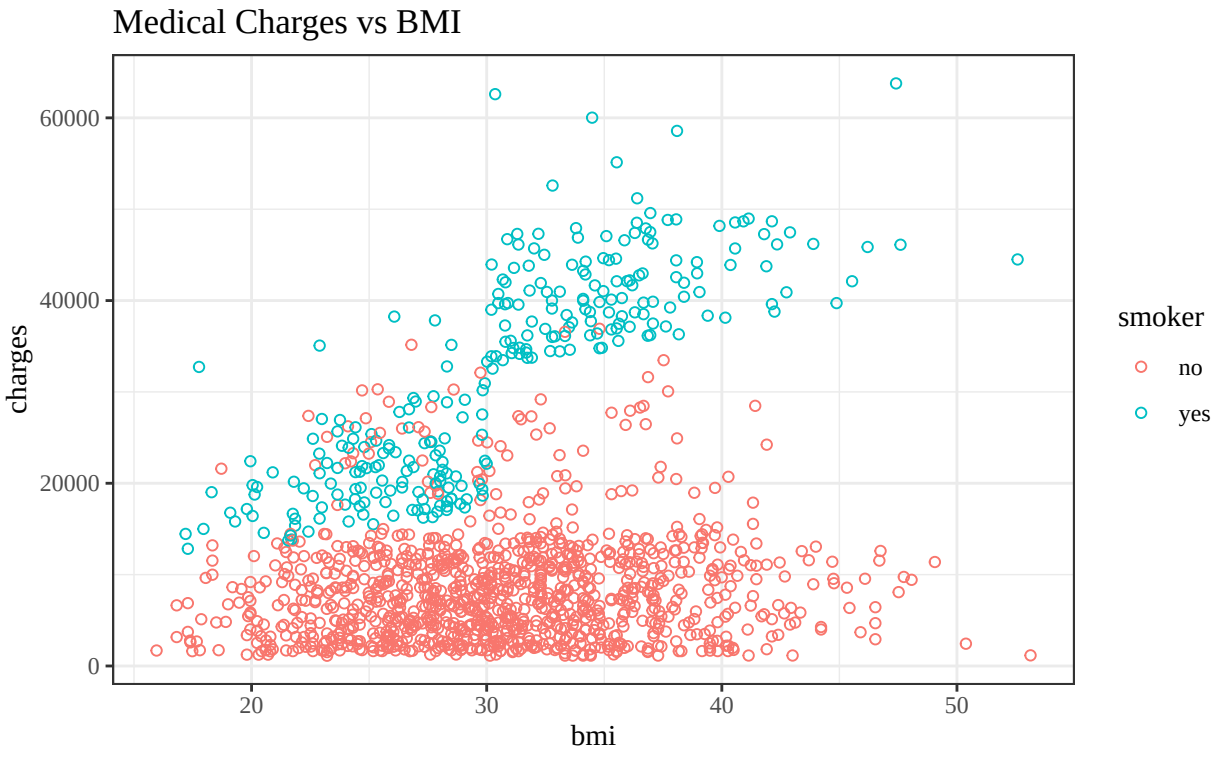
Q1

Q1.a

Table 1: Number of missing values in 'Medical Cost' data

age	sex	bmi	children	smoker	region	charges
0	0	0	0	0	0	0

Q1.b



Q1.c

Here we examine three least-squares linear models applied to the *Medical Cost* dataset to predict charges:

1. *bmi* as a sole predictor
2. *bmi* and *smoker* as predictors
3. *bmi* and *smoker* as predictors, with an interaction term

Model 1:

$$\text{charges} \sim \text{bmi}$$

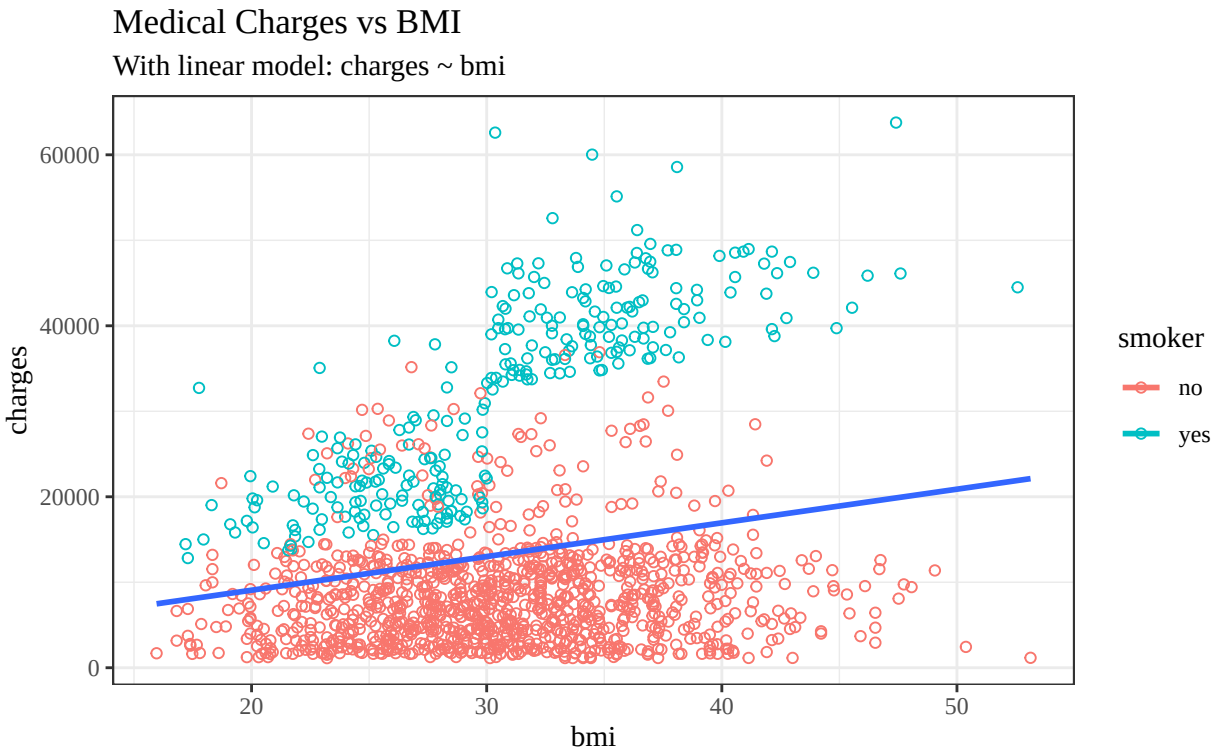


Table 2: Model 1 summary statistics

term	estimate	std. err	t-value	p-value	2.5% CI	97.5% CI
(Intercept)	1192.937	1664.802	0.717	0.474	-2072.974	4458.849
bmi	393.873	53.251	7.397	0.000	289.409	498.337

When medical cost data is fit to this model, we estimate a positive trend of $393.87 per unit of BMI. A 95% confidence interval suggests the true value is between $289.41 and $498.34.

This model has a training mean squared error (MSE) of 140777900.10.

According to this model, a smoker with a BMI of 32 would be billed $13796.87. Reducing their BMI to 28 is associated with a change in cost of -1575.49 dollars.

Model 2:

$$\text{charges} \sim \text{bmi} + \text{smoker}$$

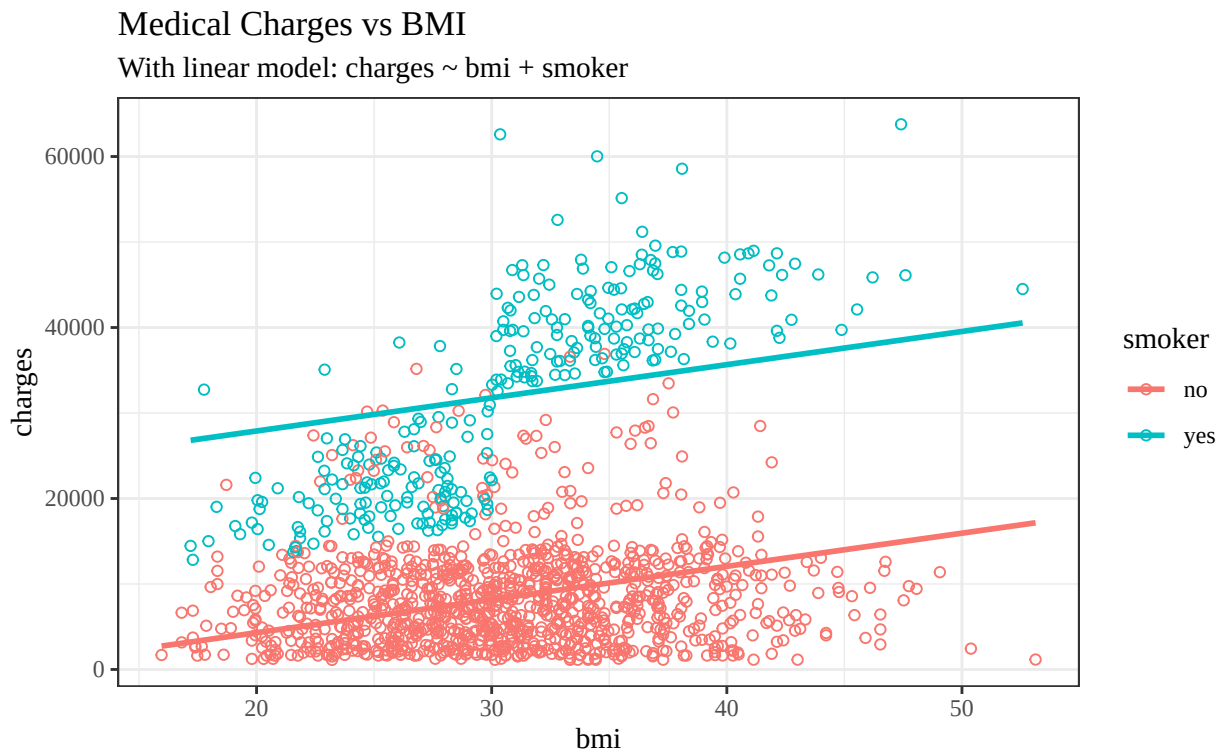


Table 3: Model 2 summary statistics

term	estimate	std. err	t-value	p-value	2.5% CI	97.5% CI
(Intercept)	-3459.096	998.279	-3.465	0.001	-5417.463	-1500.728
bmi	388.015	31.787	12.207	0.000	325.656	450.374
smokeryes	23593.981	480.180	49.136	0.000	22651.990	24535.972

When medical cost data is fit to model 2, we estimate a trend of costs changing $388.02 per unit of BMI, and a difference in cost of $23593.98 associated with being a smoker versus being a non-smoker. A 95% confidence interval suggests the true association between charge and BMI is between $325.66 and $450.37 per unit of BMI, and the true cost associated with becoming a smoker is from $22651.99 to $24535.97.

This model has a training mean squared error (MSE) of 50126126.42.

According to this model, a smoker with a BMI of 32 would be billed $32551.37. Reducing their BMI to 28 is associated with a change in cost of -1552.06 dollars.

Model 3:

$$\text{charges} \sim \text{bmi} * \text{smoker}$$

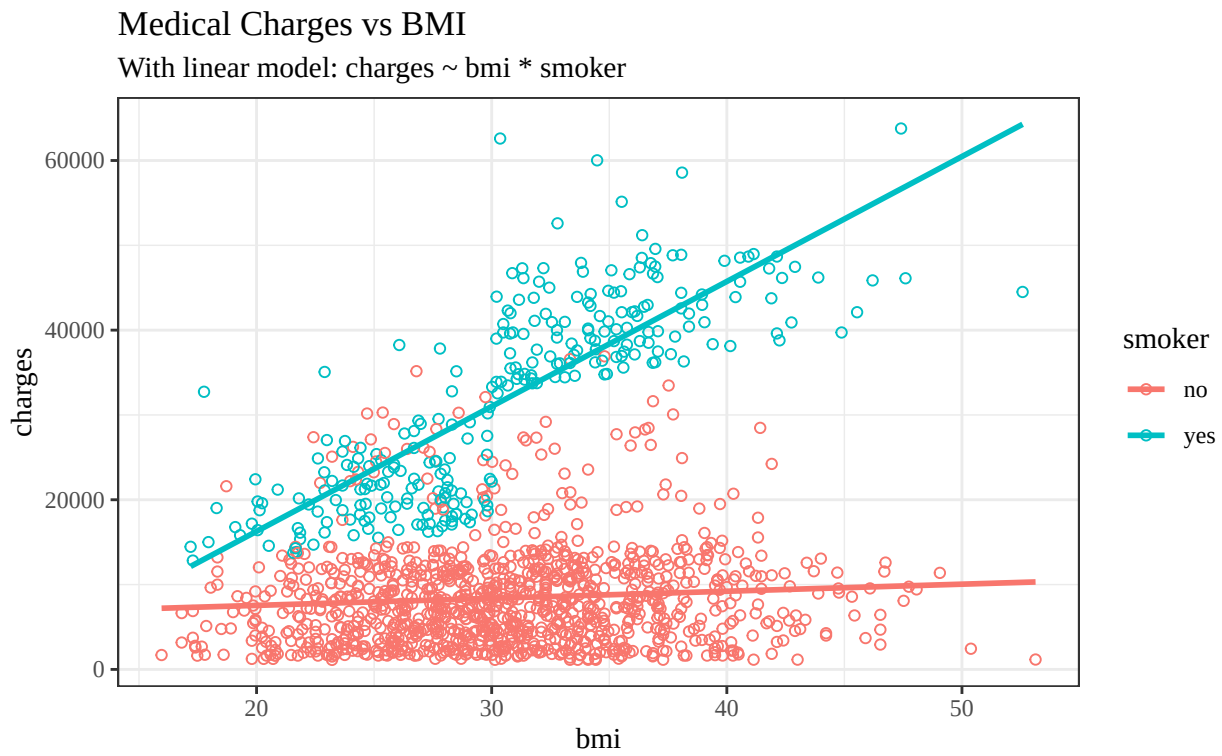


Table 4: Model 3 summary statistics

term	estimate	std. err	t-value	p-value	2.5% CI	97.5% CI
(Intercept)	5879.424	976.869	6.019	0.000	3963.057	7795.791
bmi	83.351	31.269	2.666	0.008	22.010	144.691
smokeryes	-19066.000	2092.028	-9.114	0.000	-23170.024	-14961.977
bmi:smokeryes	1389.756	66.783	20.810	0.000	1258.745	1520.767

Model 3 estimates an associated increase in medical charges of $83.35 per unit increase of BMI. For smokers, there is an *additional* increase in cost per unit BMI of $1389.76. According to the model, moving from the non-smoking to smoking group is associated with change in cost of $-19066.00. It is important to note that this value is only interpretable within the range of the provided data, so it will always be offset by observations where BMI is at least 10. In each case, the the 95% confidence interval reflect that range of values we are 95% certain the true value these estimates lie within.

This model has a training mean squared error (MSE) of 37841585.43.

According to this model, a smoker with a BMI of 32 would be billed $33952.82. Reducing their BMI to 28 is associated with a change in cost of -5892.43 dollars.

Q1.d

Adding a new variable to the data indicating if a subject is a smoker and has a BMI over 30, we can build a new model incorporating this information:

$$\text{charges} \sim \text{bmi} * (\text{smoker} + \text{smoker\_bmi30p})$$

Table 5: Model 4 summary statistics

term	estimate	std. err	t-value	p-value	2.5% CI	97.5% CI
(Intercept)	5879.424	922.775	6.371	0.000	4069.174	7689.674
bmi	83.351	29.537	2.822	0.005	25.406	141.295
smokeryes	3191.774	4230.914	0.754	0.451	-5108.207	11491.755
smoker_bmi30pTRUE	14546.031	5862.788	2.481	0.013	3044.727	26047.335
bmi:smokeryes	401.754	164.301	2.445	0.015	79.436	724.071
bmi:smoker_bmi30pTRUE	23.426	199.117	0.118	0.906	-367.191	414.043

Fitting this model to the data, the predictors with a sufficiently significant association ($p > 0.05$) to reject the null hypothesis that there is no relationship with medical charges are *bmi*, *smoker_bmi30p*, and the interaction term between *smoker* and *bmi*.

If we were to drop the non-significant terms from the model (*smoker* and the interaction between *bmi* and *smoker_bmi30p*), smokers under 30 BMI would no longer have a different estimated intercept than non-smokers, decreasing the level of charges of the smoker and smoker over 30 BMI groups. Smokers would also be associated with the same change in cost per unit BMI increase regardless of their BMI level.

Q2

Q2.a

Example 1: Estimating global fertility rates by country and time

In this case, Y is fertility rate, and our features would be proportion of females in the reproductive age group, year, region, mean years of education, and sociodemographic index. The goal of modeling this data is to produce estimates for countries with no available data. Even with some data sparse location-years, fertility is still comparatively well studied and there is much data available, making this problem low-dimensional.

Example 2: Examining cardiovascular risk factors in a cohort

Here, Y is a risk factor of interest, such as creatinine level at observation. X could include many features, but would likely include the days since observation and if the subject had died since observation, along with other potential covariates such as blood pressure level at observation, smoker status, age, sex, weight, and height. The goal of including many covariates and grouping by death is to determine which risk factors retain a significant association. While studies like this can capture many dimensions, there are usually still more participants, making this a low-dimensional problem.

Example 3: Predicting air quality level in a population

The outcome of this problem is predicted air quality level, and the input data would include features such as time of day, day of year, temperature, rain level, amount of traffic, and proximity to factories. The goal of such models could be to act as a short-term forecasting system that can be used to inform at-risk groups of the potential danger of outdoor activities. Since the inputs can be derived largely from surveillance monitoring systems, there would be an abundance of observations compared to features, making this another low-dimensional problem.

Q2.b

Example 1: Classifying malignant tumors in MRI images

The outcome is a binary value indicating the presence of a malignant tumor in an image. The data are pixels in the image, coded as grayscale values. The goal of such a model is to detect abnormalities in images that

represent a malignant tumor more reliably than manual review. This problem is high-dimensional, since each pixel in an image represents one feature.

Example 2: Genomic sequencing for gene expression

Here, the outcome is the presence of a specific gene, and the data are the genomic sequences of each sample. The goal is to detect patterns in a sequence that indicate the expression of the gene of interest. Since long sequences can be collected, this data set is wide or similar in size compared to the number of observations, making this a high dimensional-problem.

Example 3: Predicting health care access

In this example, the outcome is predicting if individuals have adequate access to health care systems. The input data would have features like income level, education, sex, age, location (urban/rural), and government healthcare spending. Using these variables, a logistic regression could be fit to get the proportion of a population with sufficient access to health care. Since this model uses few features, there will be more observations than features, making this a low-dimensional problem.

Q2.c

Example 1: Identifying disease spread in population with social media data

The outcome is a probability of there being a new outbreak spreading through a population. The features of this data are messages, timestamps, and location data made available through social media services. The goal is to analyze message content, frequency, and location to predict if disease spread is sufficiently likely. For a population with easy access to social media, there will be many observations, but the breadth of data collected is wide enough (especially for encoding messages) to make this a high-dimensional problem.

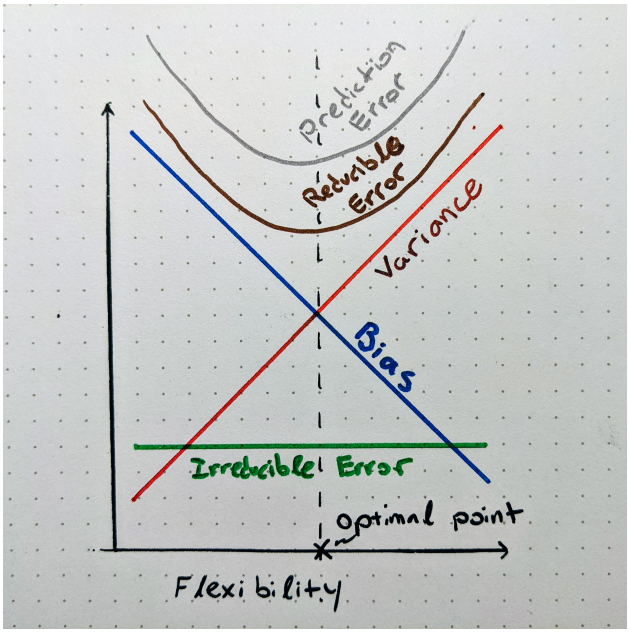
Example 2: Identifying pathogen mutations

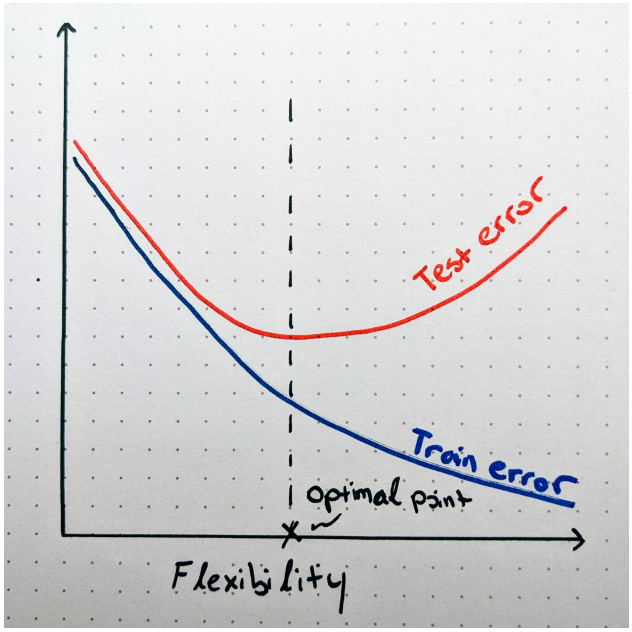
This model would determine groupings of pathogen sample to determine if mutations are present. Here, the outcome would be an assigned categorical group, which is determined by input data such as genome sequencing collected from different samples of a pathogen in a population. Since the characteristics of a mutation are not known ahead of time, this model would try to detect patterns based on similarities and differences to other observations. Collection would limit the sample size, but a lot of information can be extracted from any sample, making this a high-dimensional problem.

Example 3: Discovering novel drug treatments for a disease

In this example, the outcome is the effectiveness of a new drug for treating a disease. The input data would be the chemical composition of the drug, along with delivery method and frequency. The task is to create a reinforced learning model that will simulate and test different combinations, iterating on what works based on some loss criteria. The details of chemical composition and combination create a lot of data per observation, making this a high-dimensional problem.

Q3





A simple regression ($p = 1$) would fall to the low end of flexibility, since such a regression would vary little with different samples of the training set (low variance), and not fit any observation well (high bias). A K-nearest neighbors model with $K = 1$, on the other hand, would belong on the high end of flexibility since the model would fit observed data well (low bias), but vary greatly between different sample of the training data (high variance).

Q4

When interpretability of a model is important, linear models make more sense since they provide the contribution of each term (feature) included in the model, potentially giving you the information needed to remove insignificant features and simplify your model further. On the other hand, if only the outputs of the model are of interest, then non-parametric models can make use of more complicated feature transformations to model more complicated relationships at the expense of interpretability.

Datasets with observed correlations between features and outcomes lend themselves well to parametric models. Datasets with fewer observations are also better suited to parametric models since non-parametric models need many observations to generate accurate estimate. Finally, if there is prior knowledge about the true function relating features to the outcome, then that is best encoded in a parametric model.

Q5

Using only *region* to predict charges, we re-code *region* as dummy variables to get a model of the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

Q5.a

Table 6: Region model 0/1 encoding summary

term	estimate	std. err	t-value	p-value	2.5% CI	97.5% CI
(Intercept)	12417.575	670.263	18.526	0.000	11102.691	13732.460
regionnortheast	988.809	948.626	1.042	0.297	-872.153	2849.771
regionsoutheast	2317.836	922.156	2.513	0.012	508.803	4126.869
regionsouthwest	-70.638	947.895	-0.075	0.941	-1930.165	1788.889

First, we make the model conditional on the following variable values:

- $x_{i1} = 0$: subject not from northeast; $x_{i1} = 1$: subject from northeast
- $x_{i2} = 0$: subject not from southeast; $x_{i2} = 1$: subject from southeast
- $x_{i3} = 0$: subject not from southwest; $x_{i3} = 1$: subject from southwest

With this conditioned model, the intercept represents the mean charges associated with being from the northwest region, and the estimated coefficients of the dummy variables represent the estimated change in mean charges associated with moving from the northwest region to that respective region.

Q5.b

Table 7: Region model -0.5/+0.5 encoding summary

term	estimate	std. err	t-value	p-value	2.5% CI	97.5% CI
(Intercept)	14035.579	661.487	21.218	0.000	12737.910	15333.248
regionnortheast	988.809	948.626	1.042	0.297	-872.153	2849.771
regionsoutheast	2317.836	922.156	2.513	0.012	508.803	4126.869
regionsouthwest	-70.638	947.895	-0.075	0.941	-1930.165	1788.889

Next, we recondition the model on a different set of variable values:

- $x_{i1} = -0.5$: subject not from northeast; $x_{i1} = 0.5$: subject from northeast
- $x_{i2} = -0.5$: subject not from southeast; $x_{i2} = 0.5$: subject from southeast
- $x_{i3} = -0.5$: subject not from southwest; $x_{i3} = 0.5$: subject from southwest

Now the intercept represents the estimated overall mean charges, ignoring the effect of region, while the other coefficients represent the estimated amount charges differ from the overall mean in a region.

Appendix

```r
# Prep work ---------------------------------------------------------------

library(dplyr)
library(ggplot2)
library(purrr)
library(broom)

pred_point <- function(model, ...) predict(model, newdata = data.frame(...))


# Question 1a -------------------------------------------------------------

load("data/Medical_Cost.RData")
df <- as_tibble(df)

missing_summary <- df %>%
  summarise(across(everything(), ~sum(is.na(.x))))


# Question 1b -------------------------------------------------------------

scatter_med_cost <-
  ggplot(df, aes(x = bmi, y = charges, color = smoker)) +
  geom_point(shape = 1) +
  theme_bw(base_family = "serif") +
  labs(title = "Medical Charges vs BMI")


# Question 1c -------------------------------------------------------------

# Build models
summary_colnames <- c(
  "term", "estimate", "std. err", "t-value", "p-value", "2.5% CI", "97.5% CI"
)

model_tbl <-
  tibble(
    formula = c(charges ~ bmi, charges ~ bmi + smoker, charges ~ bmi*smoker)
  ) %>%
  mutate(
    model = map(formula, lm, data = df),
    summary = map(model, tidy, conf.int = TRUE),
    mse = map(model, ~mean(residuals(.x)^2)),
    preds = map(model, ~pred_point(.x, bmi = c(32, 28), smoker = "yes"))
  )

# Visualize models
scatter_regression_1 <- scatter_med_cost +
  geom_smooth(aes(color = NULL), method = "lm", se = FALSE) +
  labs(subtitle = "With linear model: charges ~ bmi")
```

```r
scatter_regression_2 <- scatter_med_cost +
  geom_line(mapping = aes(y = predict(model_tbl$model[[2]], df)), size = 1) +
  labs(subtitle = "With linear model: charges ~ bmi + smoker")

scatter_regression_3 <- scatter_med_cost +
  geom_smooth(method = "lm", se = FALSE) +
  labs(subtitle = "With linear model: charges ~ bmi * smoker")


# Question 1d ------------------------------------------------------------

df2 <- df %>% mutate(smoker_bmi30p = smoker == "yes" & bmi > 30)

model_4 <- lm(charges ~ bmi* (smoker + smoker_bmi30p), data = df2)
model_4_summary_tbl <- tidy(model_4, conf.int = TRUE)


# Question 5 -------------------------------------------------------------

dummy_region <- model.matrix(~region + 0, data = df)
region_formula <- charges ~ regionnortheast + regionsoutheast + regionsouthwest


# Question 5a ------------------------------------------------------------

region_summary_tbl_1 <-
  as_tibble(dummy_region) %>%
  bind_cols(df["charges"]) %>%
  lm(formula = region_formula, data = .) %>%
  tidy(conf.int = TRUE)


# Question 5b ------------------------------------------------------------

region_summary_tbl_2 <-
  as_tibble(dummy_region - .5) %>%
  bind_cols(df["charges"]) %>%
  lm(formula = region_formula, data = .) %>%
  tidy(conf.int = TRUE)
```