Business Statistics Final Exam

Fall 2018: BUS41000

This is a closed-book, closed-notes exam. You may use any calculator.

Please answer all problems in the space provided on the exam.

Read each question carefully and clearly present your answers.

Honor Code Pledge: "I pledge my honor that I have not violated the University Honor Code during this examination."

Sign: ____

Name: ____

Useful formulas

- $E(aX + bY) = aE(X) + bE(Y)$
- $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2ab \times Cov(X, Y)$
- $Cov(aX + bY, Z) = a \times Cov(X, Z) + b \times Cov(Y, Z)$
- The standard error of $\bar{X}$ is defined as $s_{\bar{X}} = \sqrt{\frac{s_X^2}{n}}$, where s_X^2 denotes the sample variance of X.
- The standard error for the difference in the averages between groups a and b is defined as:

$$s_{(\bar{X}_a - \bar{X}_b)} = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$$

  where s_a^2 denotes the sample variance of group a and n_a the number of observations in group a.
- The standard error for a proportion is defined by: $s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- The standard error for difference in proportion is defined by:

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

  where $\hat{p}_1$ and $\hat{p}_2$ denote two independent proportions, and n_1 and n_2 are the number of trials.
- Bayes formula:

$$P(A|B) = \frac{P(A \text{ } and \text{ } B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

  where A, B are two events.
- For $Z \sim N(0, 1)$, $P(-1 \leq Z \leq 1) = 68\%, P(-2 \leq Z \leq 2) = 95\%, P(-3 \leq Z \leq 3) = 99\%$.
- Similarly, $X \sim N(\mu, \sigma^2)$, $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 95\%$.
- Standardization to standard normal: assume $X \sim N(\mu, \sigma^2)$, $Z \sim N(0, 1)$, then

$$P(a \leq X \leq b) = P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}).$$

- Correlation: $Cor(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) \times Var(Y)}}$.

Problem 1: Fraud Detection

Credit-card fraud costs business in the United States billions of dollars each year in stolen goods. Compounding the problem, the risk of fraud increases with the rapidly growing online retail market. To reduce fraud, businesses and credit card vendors have devised systems that recognize characteristics of fraudulent transactions. These systems are not perfect, however, and sometimes flag honest transactions as fraudulent and sometimes miss fraudulent transactions.

A business has been offered a fraud detection system to protect its online retail site. The system promises very high accuracy. The system catches 99% of fraudulent transactions; that is, *given* a transaction is fraudulent, the system signals a problem 99% of the time. The system flags honest transactions as fraudulent only 2% of the time.

1. The description of this system gives several conditional probabilities, but are there other conditional probabilities that are more relevant to owners of the retail site? [4 points]

$$P(\text{ fraudulent } \mid \text{ positive })$$

2. Suppose that the prevalence of fraud among transactions at the retailer is 1%. What are the chances that the transactions are honest *and* the system incorrectly labels them as fraud? [4 points]

$$P(\text{ honest } \textit{and} \text{ positive }) = P(\text{ positive} \mid \text{ honest }) \times P(\text{ honest }) = 0.02 \times 0.99 = 0.0198$$

3. Suppose that the prevalence of fraud among transactions at the retailer is 1%. What is the probability that the transaction is an actual fraudulent, *given* the system classified it as a fraud? What would your answer be if the prevalence of fraud is higher, say 5%? [4 points]

$$P(\text{ fraud } \mid \text{ positive }) = \frac{P(\text{ positive } \textit{and} \text{ fraud })}{P(\text{ positive } \textit{and} \text{ fraud }) + P(\text{ positive } \textit{and} \text{ honest })} = \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.02 \times (1 - 0.01)} = \frac{1}{3}$$

If 0.05,

$$P(\text{ fraud } \mid \text{ positive }) = \frac{P(\text{ positive } \textit{and} \text{ fraud })}{P(\text{ positive } \textit{and} \text{ fraud }) + P(\text{ positive } \textit{and} \text{ honest })} = \frac{0.99 \times 0.05}{0.99 \times 0.05 + 0.02 \times (1 - 0.05)} = 0.723$$

4. Suppose the probability that the transaction is an actual fraudulent given the system classified it as a fraud, is 50%. In this case, what is the prevalence level of fraud among transactions at the retailer? [4 points]

We need to solve the equation below

$$0.5 = \frac{0.99 \times x}{0.99 \times x + 0.02 \times 0.99(1 - x)}$$

So $x = 0.0198$

5. There are two kinds of errors in this system. We call the error to be “Type I error”, when the transactions are honest *and* the system incorrectly labels them as fraud. And “Type II error”, when the transactions are fraudulent *and* the system let them pass as honest ones. Suppose that the prevalence of fraud among transactions at the retailer is 1%. *Given* the fact that there is an error, what are the chances that it is a Type I error? [4 points]

$$\begin{aligned} P(\text{ Type 1 } \mid \text{ error }) &= \frac{P(\text{ Type 1 } \textit{and} \text{ error })}{P(\text{ Type 1 } \textit{and} \text{ error } + P(\text{ Type 2 } \textit{and} \text{ error })} \\ &= \frac{P(\text{ positive } \textit{and} \text{ honest })}{P(\text{ positive } \textit{and} \text{ honest }) + P(\text{not positive } \textit{and} \text{ fraud })} \\ &= \frac{0.02 * 0.99}{0.02 * 0.99 + 0.01 * 0.01} = 0.995 \end{aligned} \tag{1}$$

Problem 2: Portfolios

Consider we are building portfolios from three stocks: Apple, Safeway and Intel. We collect 67 months of data to estimate the mean and standard deviation of the monthly returns, for each stock:

Stock	Sample Mean	Sample Standard Deviation
Apple	3.3	7.5
Safeway	1.4	8.3
Intel	1.9	6.5

Assume that we model the monthly returns by normal distribution.

1. Based on these statistics, what are the Sharpe ratios of these stocks? Which one has the highest Sharpe ratio? [4 points]

$$\text{Apple } = \frac{3.3}{7.5} = 0.44$$
$$\text{Safeway } = \frac{1.4}{8.3} = 0.17$$
$$\text{Intel } = \frac{1.9}{6.5} = 0.29$$

2. Assume monthly return follows a normal distribution, say $Apple \sim N(\mu_A, \sigma_A^2)$. Can you build a 68% confidence interval for μ_A? What is the 68% confidence interval for the true Sharpe ratio μ_A/σ_A? [4 points]

μ_{Apple} : $[\hat{\mu}_{\text{Apple}} - 1.0 \times \text{s.e.}, [\hat{\mu}_{\text{Apple}} + 1.0 \times \text{s.e.}] = [3.3 - \frac{7.5}{\sqrt{67}}, 3.3 + \frac{7.5}{\sqrt{67}}] = [2.38, 4.21]$

Sharp ratio : $[\frac{2.38}{7.5}, \frac{4.22}{7.5}] = [0.32, 0.56]$

3. One is considering to hedge the risk by holding Intel and Safeway together.

$$Portfolio1 = 0.5Intel + 0.5Safeway.$$

Assume the $Cov(Intel, Safeway) = 10$. What is the Sharpe ratio of Portfolio1? [4 points]

$$E(\mu) = 0.5 * 1.9 + 0.5 * 1.4 = 1.65$$
$$SD(\mu) = \sqrt{(0.5 * 6.5)^2 + (0.5 * 8.3)^2 + 2 * 0.5 * 0.5 * 10} = 5.73$$

The Sharp ratio is $\frac{1.65}{5.73} = 0.29$

4. Consider one tries to improve the Portfolio1 by considering a Portfolio2 which combines three stocks

$$Portfolio2 = 0.6 Apple + 0.4 Portfolio1.$$

Assume that $Cov(Apple, Intel) = 19$, $Cov(Apple, Safeway) = 0.002$.

- Can you show me the derivation steps, why $Cov(Apple, Portfolio1) = 9.501$? [2 points]

$$Cov(A, P) = Cov(A, 0.5I + 0.5S) = 0.5Cov(A, I) + 0.5Cov(A, S) = 9.501$$

- What is the Sharpe Ratio of Portfolio2? [2 points]

$$\mu = 0.6 * 3.3 + 0.4 * 1.65 = 2.64$$
$$SD = \sqrt{0.6^2 * Var(A) + 0.4^2 * Var(P1) + 2 * 0.6 * 0.4 * Cov(A, P1)} = 5.48$$
$$Sharpratio = \frac{2.64}{5.48} = 0.48$$

5. Now consider both Portfolio1 and Portfolio2, can you build 95% confidence intervals for the true Sharpe ratio of both? Do these two intervals overlap? Can you say with confidence that Portfolio2 is better than Portfolio1 in terms of Sharpe ratio, based on only these 67 samples? [4 points]

Confidence interval of average return

$$\mu_1 : [1.65 - 2\frac{5.73}{\sqrt{67}}, 1.65 + 2\frac{5.73}{\sqrt{67}}] = [0.24, 3.06]$$
$$\mu_2 : [2.64 - 2\frac{5.48}{\sqrt{67}}, 2.64 + 2\frac{5.48}{\sqrt{67}} = [1.3, 3.98]$$

Confidence interval of Sharp ratio

$$S_1 : [\frac{0.24}{5.73}, \frac{3.06}{5.73}] = [0.04, 0.53]$$
$$S_2 : [\frac{1.41}{4.98}, \frac{3.87}{4.98}] = [0.24, 0.73]$$

Problem 3: SLR and Inference

Consider the SLR model $Y = \beta_0 + \beta_1 X + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. Assume we go ahead collect 100 data ($n = 100$), and fit the least square line, we get

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8535     0.6137   1.391    0.167
x             1.9986     0.1436  13.914   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.851 on 98 degrees of freedom
Multiple R-squared:  0.6639,    Adjusted R-squared:  0.6605
F-statistic: 193.6 (on 1 and 98 DF),  p-value: < 2.2e-16
```

1. What is the estimate of intercept $\widehat{\beta}_0$ and slope $\widehat{\beta}_1$ based on the data? What is the correlation between X and Y? [4 points]

$\hat{\beta}_0 = 0.8535$

$\hat{\beta}_1 = 1.9986$

$corr(X, Y) = \sqrt{R^2} = 0.81$

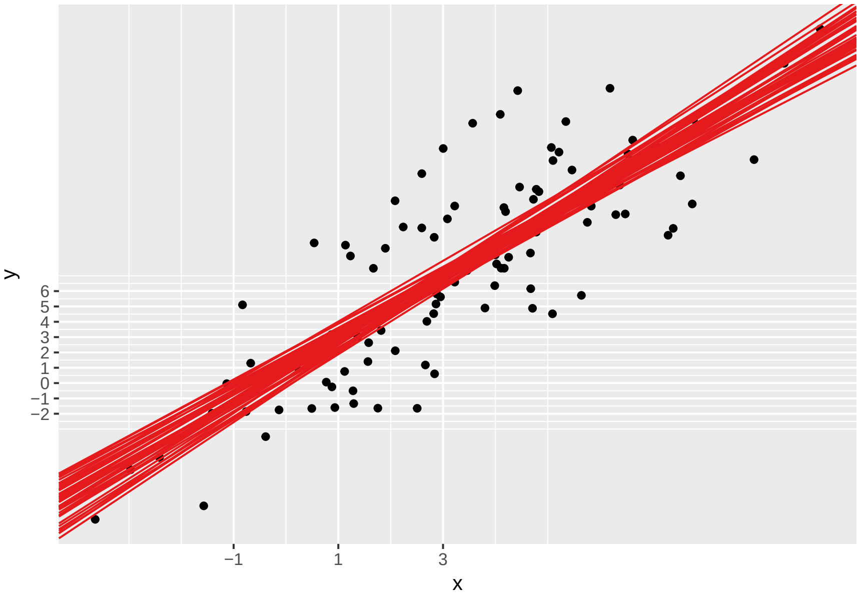
2. What is the variance of $\widehat{\beta}_1$? Can you build a 95% confidence interval for β_1? Based on the data, can you reject the null $H_0 : \beta_1 = 1$? [4 points]

$var(\hat{\beta}_1) = se(\hat{\beta}_1)^2 = 0.00206$

$CI = [1.9986 - 2 \times 0.1436, 1.9986 + 2 \times 0.1436] = [1.7114, 2.2858]$, reject $H_0 : \beta_1 = 1$

3. I asked 40 students to each independently collect 100 data points, fit their regression line, and report to me. Then I plotted all 40 regression line in red. Based on the below plot, can you tell me the sampling variance of the intercept $\widehat{\beta}_0$? Does it agree with the regression table?

$\widehat{var}(\hat{\beta}_0) = (2.5/4)^2 = 0.39$, while in the table it's 0.38, so they are close.



4. What is the residual standard error? What is the *prediction interval* for $Y|X = 1$? [2 points]

se = 3.851

$95\% PI = [2.8521 - 2 \times 3.851, 2.8521 + 2 \times 3.851] = [-4.85, 10.55]$

5. You are curious about building a **confidence interval** for $\mathbb{E}[Y|X = 1] = \beta_0 + \beta_1$. You recalled that one can use the least squares predicted value $\widehat{\beta}_0 + \widehat{\beta}_1$ as an estimate for the true $\beta_0 + \beta_1$.

- Suppose that I told you $Cov(\widehat{\beta}_0, \widehat{\beta}_1) = -0.0686$, can you calculate $Var(\widehat{\beta}_0 + \widehat{\beta}_1)$? What is the standard error of the estimate $\widehat{\beta}_0 + \widehat{\beta}_1$? (Hint, you can roughly double check your answer with the above plot) [4 points]

$var(\hat{\beta}_0 + \hat{\beta}_1) = 0.26$

$se(\hat{\beta}_0 + \hat{\beta}_1) = 0.51$

- What is the 95% confidence interval for $\mathbb{E}[Y|X = 1] = \beta_0 + \beta_1$? [2 points]

$95\% CI = [2.8521 - 2 \times 0.51, 2.8521 + 2 \times 0.51] = [1.83, 3.87]$

- Can you explain why the confidence interval for $\mathbb{E}[Y|X = 10]$ is much wider than $\mathbb{E}[Y|X = 3]$? [Bonus, 3 points]

$var(\hat{E}[Y|X = 10]) = var(\hat{\beta}_0) + 100var(\hat{\beta}_1) + 20cov(\hat{\beta}_1, \hat{\beta}_0)$

$var(\hat{E}[Y|X = 3]) = var(\hat{\beta}_0) + 9var(\hat{\beta}_1) + 6cov(\hat{\beta}_1, \hat{\beta}_0)$

Problem 4: MLR on Boston Housing

In this question, we take a look at a data-set that contains information collected by the U.S Census Service concerning housing in the area of Boston Mass.

variables	
medv	median value of owner-occupied homes in \$1000s
rm	average number of rooms per dwelling
dis	weighted mean of distances to five Boston employment centres
lstat	lower status of the population (percent)

We first work with the following MLR model

$$\text{Model 1:} \quad medv = \beta_0 + \beta_1 \cdot rm + \beta_2 \cdot dis + \epsilon.$$

Here is the MLR output.

```
Model 1
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -34.6361     2.6215 -13.212  < 2e-16 ***
rm            8.8014     0.4236  20.780  < 2e-16 ***
dis           0.4888     0.1413   3.459 0.000588 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.545 on 503 degrees of freedom
Multiple R-squared:  0.4955,    Adjusted R-squared:  0.4935
F-statistic:   247 (on 2 and 503 DF),  p-value: < 2.2e-16
```

1. Provide an interpretation for the coefficients associated with **rm** and **dis**? [4 points]

rm: Given **dis** fixed, one unit increase in **rm** is associated with 8.8014 more units of **medv** on average.

dis: Given **rm** fixed, one unit increase in **dis** is associated with 0.4888 more unit of **medv** on average.

2. What is the F-statistic telling you? Clearly explain the hypothesis being tested and your conclusion. [4 points]

At least one of **rm** and **dis** matters

$H_0 : \beta_1 = \beta_2 = 0$

3. What is the t-statistic for **dis** telling you? Clearly explain the hypothesis being tested and your conclusion. [2 points]

dis’s association with **medv** is significantly different from zero

$H_0 : \beta_2 = 0$

In fitted Model 1, as distance to employment center increase (controlling for rooms fixed), the median house prices increase. This is a bit hard to believe, as houses tend to be more expansive if the commuting time is shorter. We will investigate this matter using the following MLR model.

$$\text{Model 2:} \quad medv = \beta_0 + \beta_1 \cdot rm + \beta_2 \cdot dis + \beta_3 \cdot lstat + \epsilon.$$

```
Model 2
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.80829    3.36800   0.834 0.404781
lstat       -0.72333    0.04933 -14.662  < 2e-16 ***
rm           4.87339    0.44456  10.962  < 2e-16 ***
dis         -0.46128    0.13495  -3.418 0.000682 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.482 on 502 degrees of freedom
Multiple R-squared:  0.6468,    Adjusted R-squared:  0.6447
F-statistic: 306.4 (on 3 and 502 DF),  p-value: < 2.2e-16
```

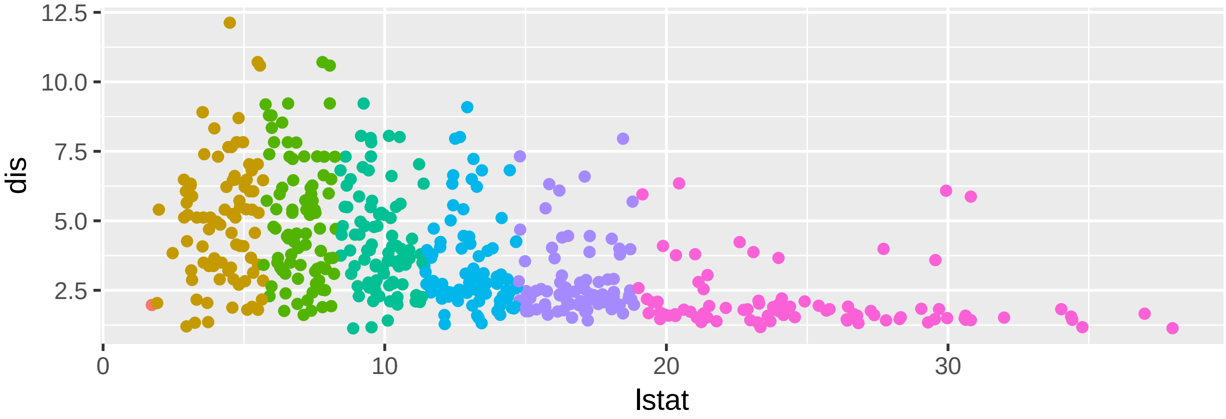
4. Compare to Model 1, is Model 2 better? State which numbers you are using from the regression table to draw your conclusion. In addition, how to interpret the coefficients associated with **dis** now? [4 points]

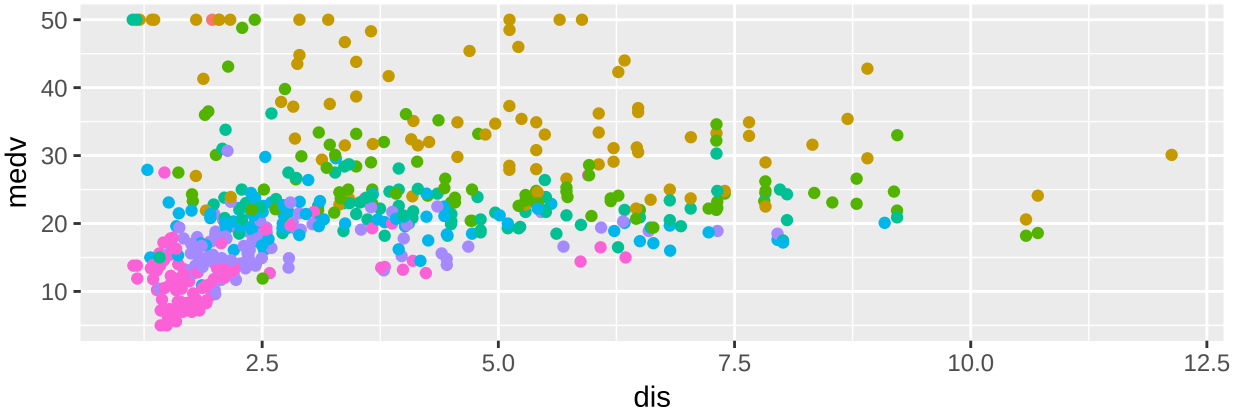
Yes, since Model 2 has lower residual standard error.

Given **rm** and **lstat** fixed, one unit increase in **dis** is associated with 3.418 less units of **medv** on average.

5. Let’s try to explain why there is a sign change in the slope of **dis** for Model 1 and Model 2. Based on the below two plots, choose the option where the correlation make most sense. (Hint: you can ignore the color for now) [2 points]

- [A] $Cor(dis, lstat) = -0.497$, $Cor(medv, dis) = 0.250$ - Correct
- [B] $Cor(dis, lstat) = 0.497$, $Cor(mdev, dis) = -0.250$
- [C] $Cor(dis, lstat) = 0.497$, $Cor(medv, dis) = 0.250$
- [D] $Cor(dis, lstat) = -0.497$, $Cor(medv, dis) = -0.250$





6. In the above plot, I color coded 6 sub-groups of the data, where each color/sub-group roughly corresponds to holding *lstat* "constant". Choose the options that make sense, combining all the information above. (Hint: you can choose multiple ones) [4 points]

- [A] The above color plot shows when holding percentage of lower status population fixed, houses that are far away from the Boston employment centres tend to be less expensive than houses that are close, on average.
- [B] The above color plot shows that for houses with very high percentage of lower status population, it is most likely they are very close to Boston employment centres.
- [C] The above color plot shows that for houses that are very far way from the Boston employment centres, it is most likely they have very low percentage of lower status population.
- [D] The positive coefficient of **dis** in Model 1 is due to the fact that variables **dis** and **lstat** are negatively correlated, meaning that a larger distance usually correlates with lower percentage of lower status population, which ends up with a higher price on average.

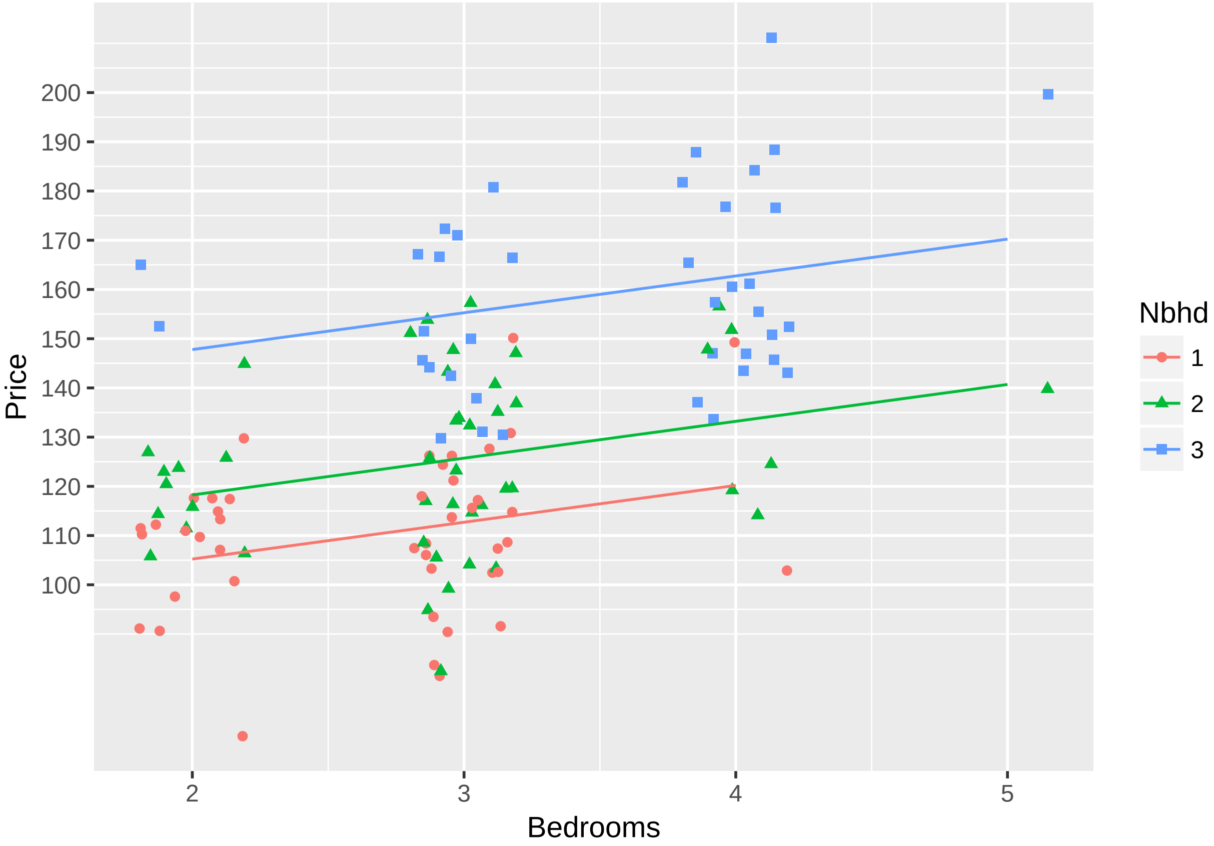
All Correct!

Problem 5: Dummies and Interactions

Let's go back to the Midcity housing prices data-set from our homework. Let's start by looking at the following model

$$\text{Model 1:} \quad Price = \beta_0 + \beta_1 \cdot Nbhd2 + \beta_2 \cdot Nbhd3 + \beta_3 \cdot Bedrooms + \epsilon$$

where $Nbhd2, Nbhd3$ are dummies for neighborhood 2 and 3 respectively, and $Bedrooms$ is **not** a dummy.



Based on the above regression plot for Model 1, answer the following questions. Here the legend 1 means for houses in neighborhood 1.

1. What is the estimated value for the effect of Bedrooms on Prices for houses β_3? What is the the estimated value for β_0? [4 points]

- [A] $\beta_3 = 7.48$, $\beta_0 = 105$
- [B] $\beta_3 = 5.98$, $\beta_0 = 105$
- [C] $\beta_3 = 7.48$, $\beta_0 = 90$ - Correct

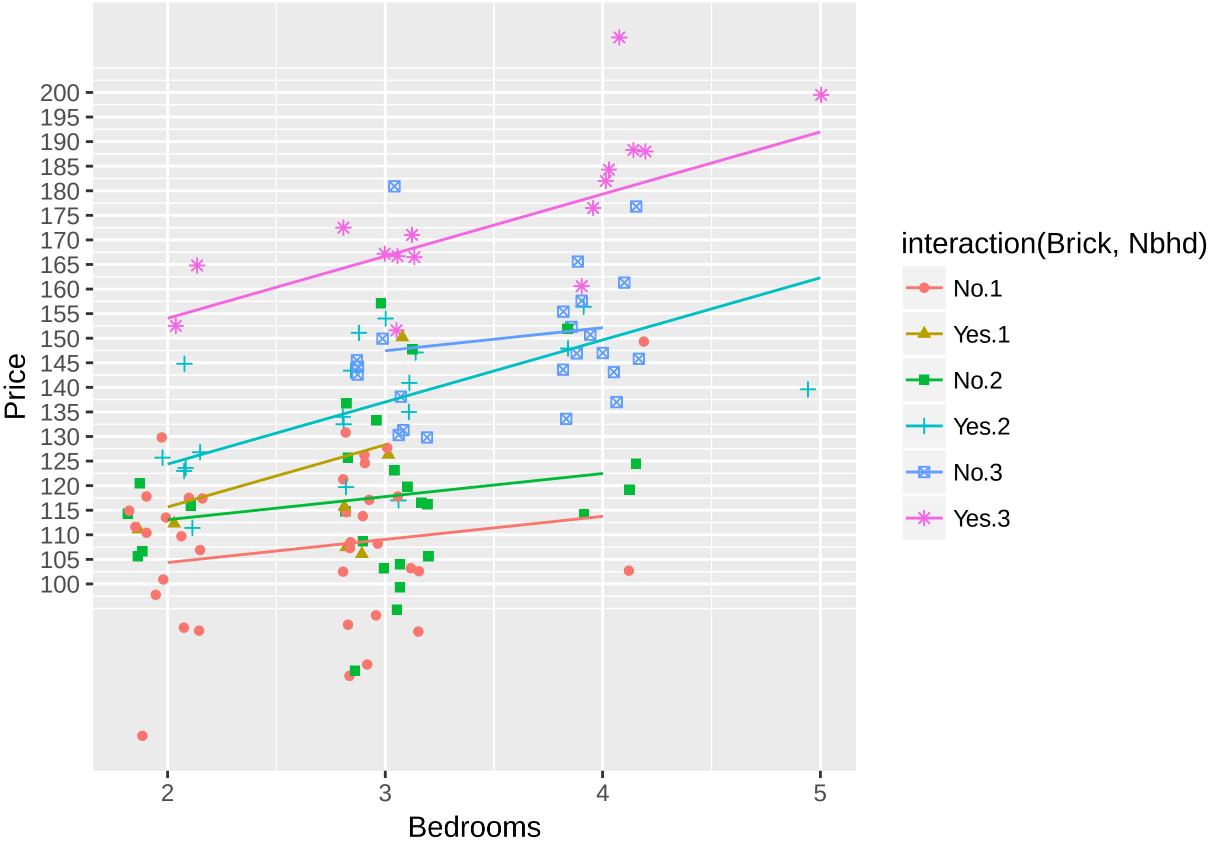
2. What is the estimated average difference (call it Δ_1) between 2-bedrooms house in neighborhood 2 and neighborhood 1? What is the estimated average difference (call it Δ_2) between a 3-bedrooms house in neighborhood 3 and neighborhood 1? What is the residual standard error s? [4 points]

- [A] $\Delta_1 = 13.0$, $\Delta_2 = 42.6$, $s = 17.29$ - Correct
- [B] $\Delta_1 = 13.0$, $\Delta_2 = 29.6$, $s = 17.29$
- [C] $\Delta_1 = -13.0$, $\Delta_2 = 29.6$, $s = 34.58$

Let's continue by looking at the following more complicated model

Model 2: $Price = \beta_0 + \beta_1 \cdot Nbhd2 + \beta_2 \cdot Nbhd3 + \beta_3 \cdot Bedrooms + \beta_4 \cdot Brick + \beta_5 \cdot Brick \times Bedrooms + \epsilon$

where *Brick* is a dummy.



Based on the above regression plot for Model 2, answer the following questions. Here the legend $No.1$ means for houses in neighborhood 1 that are build by wood, and $Yes.3$ means for house in neighborhood 3 that are build by brick, etc.

3. What is the estimated average difference between a 2-bedrooms wood house in neighborhood 1 (call it P_1), and a 4-bedrooms brick house in neighborhood 3 (call it P_2)? Choose the formula and the estimated value. [4 points]

- [A] formula $P_1 - P_2 = -\beta_2 - 2\beta_3 - \beta_4 - 4\beta_5$, estimated value -75 - Correct
- [B] formula $P_1 - P_2 = -\beta_2 - 2\beta_3 - 4\beta_5$, estimated value -80
- [C] formula $P_1 - P_2 = -\beta_2 - 2\beta_3 - \beta_4$, estimated value -45

4. What is the estimated average difference between a 3-bedrooms brick house in neighborhood 2 (call it P_3), and a 3-bedrooms wood house in the same neighborhood (call it P_4)? Choose the formula and the estimated value. [4 points]

- [A] formula $P_3 - P_4 = \beta_4 + 3\beta_5$, estimated value 19.2 - Correct
- [B] formula $P_3 - P_4 = 3\beta_3 + \beta_4$, estimated value 9.5
- [C] formula $P_3 - P_4 = \beta_4$, estimated value 19.2

5. What are the estimated sign of the estimated value for β_4, and β_5? What is the ordering of β_1 and β_2? [4 points]

- [A] $\beta_4 > 0, \beta_5 > 0,\ \beta_1 < \beta_2$
- [B] $\beta_4 < 0, \beta_5 > 0,\ \beta_1 < \beta_2$ - Correct
- [C] $\beta_4 > 0, \beta_5 > 0,\ \beta_1 > \beta_2$
- [D] $\beta_4 < 0, \beta_5 > 0,\ \beta_1 > \beta_2$

6. Is the Model 2 equibalent to running two seperate regressions on sub-population? [2 points]

- For the **Brick** houses, run a regression

$$Price = \beta_0 + \beta_1 \cdot Nbhd2 + \beta_2 \cdot Nbhd3 + \beta_3 \cdot Bedrooms + \epsilon$$

- For the **Non-Brick** houses, run a regression

$$Price = \beta_0' + \beta_1' \cdot Nbhd2 + \beta_2' \cdot Nbhd3 + \beta_3' \cdot Bedrooms + \epsilon$$

Choose [**YES**] or [**NO**].

No!