7 Question

Suppose you are in charge of making a part are you are confident that the part making process is "under control" in that whether or not a part is defective is iid Bernoulli with a 1 representing a defect and a 0 representing a good one.

So, your model is $Y_i \sim$ Bernoulli(p) iid.

However, you still need to learn about p where p is the probability that a part is defective.

7.1

Suppose you collect data on 1,000 parts and find that 95 of them are defective.
What is your estimate of p?

$$\hat{p} = \frac{95}{1000} = \underline{\underline{.095}}$$

7.2

Give a 95% confidence interval for p.

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.095(1-.095)}{1000}} \simeq .0093$$

$$\Rightarrow \quad \hat{p} \pm 2 \times SE \quad \Rightarrow \quad \underline{\underline{.095 \pm .0186}}$$

7.3

Suppose your boss claims that $p = .15$.

Using this data, test the null hypothesis $H_0 : p = .15$.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.095 - .15}{\sqrt{\frac{.15(1-.15)}{1000}}} = \underline{\underline{-4.87}}$$

$$\Rightarrow \underline{\underline{\text{reject } H_0}}$$

7.4

Do you think you have evidence to refute your boss's claim?

<u>yes! clear reject!</u>

8 Question

A company wishes to asses the effectiveness of a one day training program.

To make as assessment 100 members of the sales force were randomly selected and then 50 were randomly selected to take the training while the remaining 50 did not.

For each of the salespersons, number of units sold was collected for the week prior to the training and for the week after the training.

So, for each of 100 salespersons we have:
wk1: sales prior to training
wk2: sales after training
T: 1 if they got the training, 0 if they did not.
To analysize the data, the following multiple regression model was fit:

$$wk2 = \alpha + \beta_1 \, wk1 + \beta_2 \, T + \epsilon.$$

Here is the regression ouput:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.47174    7.02630   2.629  0.00996 **
wk1          0.81866    0.06943  11.791  < 2e-16 ***
T           -0.84577    1.47809  -0.572  0.56851
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.311 on 97 degrees of freedom
Multiple R-squared: 0.5983,Adjusted R-squared:  0.59
F-statistic: 72.22 on 2 and 97 DF,  p-value: < 2.2e-16
```

So, for example, the estimate of β_2 is -.84577 and the associated standard error, t-statistic, and p-value (for $H_0 : \beta_2 = 0$) are 1.47809, -.572, and .56851 respectively.

8.1

Given the sales in wk1 is 95, give the 95% plug-in predictive interval for sales in wk2 for someone who had the training.

$18.47 + .82 \times 95 + (-.85)(1) = \underline{\underline{95.4 \pm 2 \times 7.311}}$

$= [80.78;\ 110.02]$

8.2

Give the 95% confidence interval for β_1.

$.81866 \pm 2 \times .07 = \underline{\underline{[.68;\ .96]}}$

8.3

Test the null hypothesis $\beta_1 = 0$.

$|t\text{-stat}| < 2 \Rightarrow$ $\underline{\underline{\text{fail to reject}}}$

8.4

Give the 95% confidence interval for β_2.

$-.84577 \pm 2 \times 1.478 = [-3.80177;\ 2.11023]$

8.5

Test the null hypothesis $\beta_2 = 0$.

$\underline{\underline{\text{fail to reject}}}$

8.6

A manager looks at the ouput and comments “wow, I know how to interpret the coefficient of a dummy”, this says the effect of the training is actually negative, that’s interesting!

Is this a reasonable reaction to the ouput?

see sol!

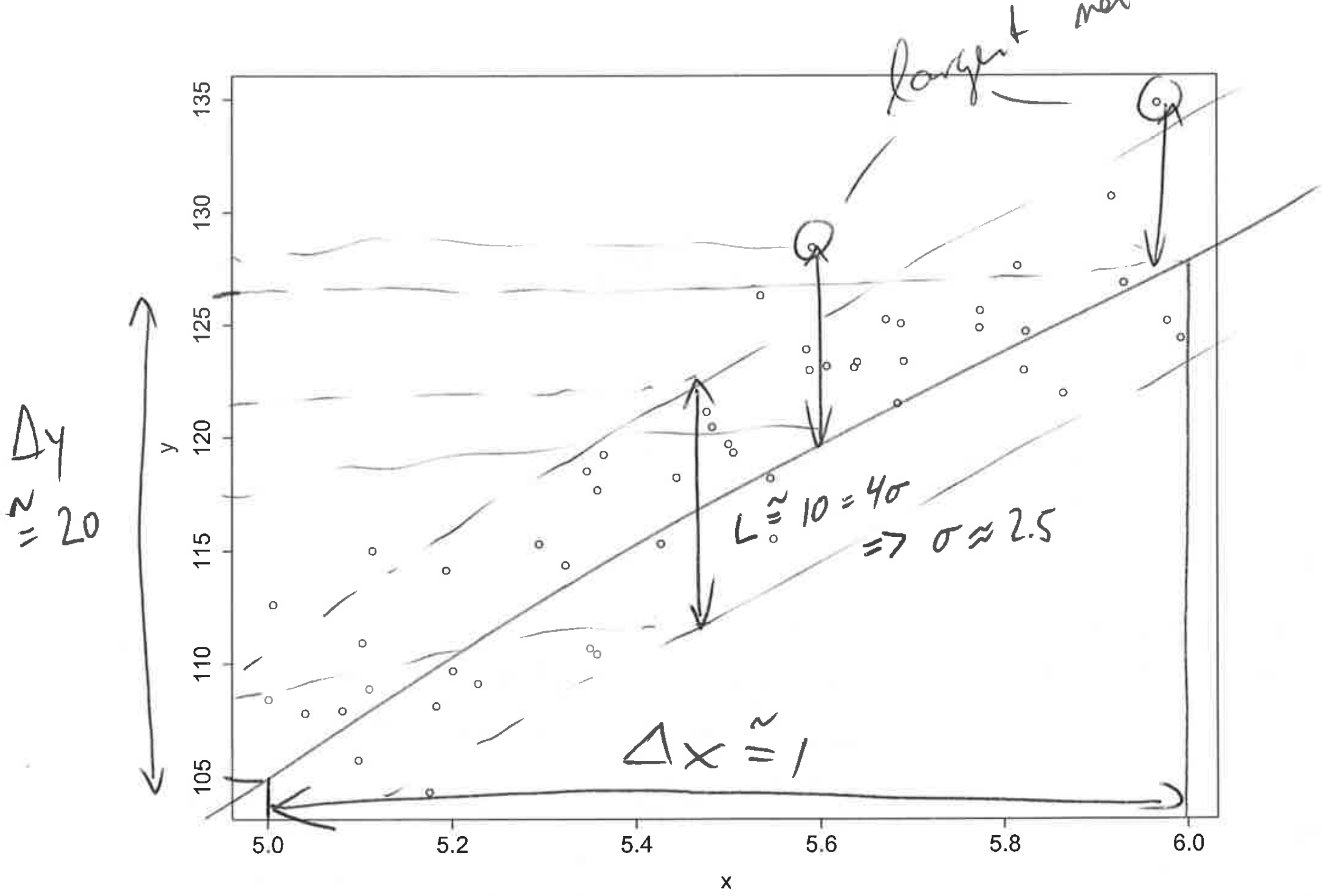
8.7

Test the null hypothesis $\beta_1 = 1$.

see sol!

9 Question

$b = \frac{\Delta y}{\Delta x} \cong \frac{20}{1} \cong 20$



The following questions refer to the x and y graphed above.
Circle your choice.

9.1

The correlation between x and y is
(a) -.78 (b) .96 (c) .23 [✓ (d) .89]

9.2

The least squares estimate of the intercept is
(a) 25 (b) 105 [✓ (c) -5] (d) 56

$105 - 5 \times b = 105 - 5 \times 22.5 \cong -7.5$

9.3

see top of previous page

The least squares estimate of the slope is
[✓ (a) 22.5] (b) 5.8 (c) -12.4 (d) 56.0

9.4

The least squares estimate of σ is
(a) .15 [✓ (b) 3.4] (c) 11.3 (d) .01

see graph

9.5

R^2 is
(a) .56 (b) .89 (c) .98 [✓ (d) .79]

$\rho^2 = .89^2 = .79$

9.6

The p-value for testing whether the intercept is equal to 0 is
(a) .001 [✓ (b) .59] (c) -.34 (d) .02

intercept is close to 0
and not measured that
precisely => fail to reject
=> large p-val

9.7

The largest residual is
[✓ (a) 7.5] (b) .038 (c) 523.9 (d) 26.7

see graph

9.8

The regression plug-in prediction for y given $x = 5.6$ is
(a) 110 (b) -5 (c) 22.53 [✓ (d) 121.1]

$$a + bX = -5 + 22.5 \times 5.6 \cong \underline{\underline{121}}$$

10 Question

(1) If the probability of a defect is p, and parts are iid, the the probability of getting 20 good parts in a row is $(1-p)^{20}$.

[✓ T] F

(2) The cdf of the standard normal distribution evaluated at -1 is pretty close to .32.

T [✓ F]

$P(Z < -1) \cong 16\%$

(3) The pdf of the uniform distribution on (-1,1) evaluated at .5 is .5.

[✓ T] F

$\text{pdf}(x) = \frac{1}{b-a} = \frac{1}{1-(-1)} = \frac{1}{2}$

(4) To include a categorical independent variable having k possible levels in a multiple regression, we use k dummy variables.

T [✓ F]

Use $k-1$ (we have the intercept)

(5) If the p-value is very small, the confidence interval must be small as well.

T [✓ F]

(6) In the regression model we assume ϵ is independent of Y.

T [✓ F]

(7) If we toss a fair coin 100 times, we would not be very surprised to get .58 for the fraction of heads.

[✓ T] F

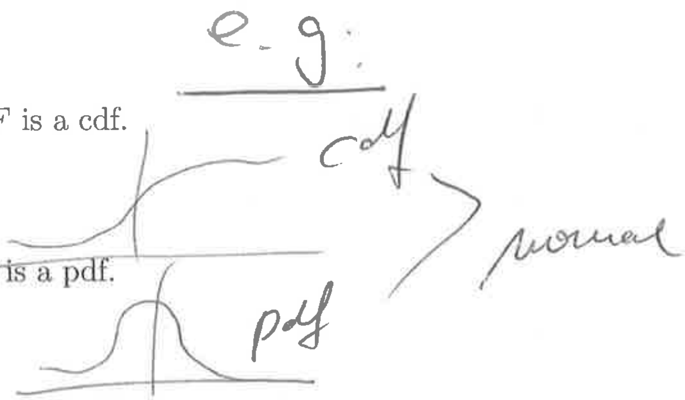
$$SE = \sqrt{\frac{.5(1-.5)}{100}} = .05$$

$$95\% \ CI = .5 \pm 2 \times .05$$

$$= [.4, .6]$$

(8) If $x_1 < x_2$ then $F(x_1) \leq F(x_2)$ where F is a cdf.

[✓ T] F



(9) If $x_1 < x_2$ then $f(x_1) \leq f(x_2)$ where f is a pdf.

T [✓ F]

(10) If $X \sim N(4, 4)$, then $Y = -4X + 10 \sim N(-6, 64)$.

[✓ T] F

$E(Y) = -4 \times 4 + 10 = -6 \quad V(Y) = (-4)^2 \cdot 4 = 64$

(11) If Y_i are iid $N(\mu, \sigma^2)$ then the probability the next 10 Y are greater than μ is $.5^{10}$

[✓ T] F

$P(Y > \mu) = .5$

(12) $cor(x, y) = cor(ax + b, y)$ where $a > 0$ and b are constants.

[✓ T] F

(13) The correlation and covariance always have the same sign.

[✓ T] F

, because $\sigma \geq 0$ always.

(14) In multiple regression when you add an explanatory variable (an x) to the regression, R^2 cannot go down.

[✓ T] F

(15) The residuals for a least squares line sum to zero.

[✓ T] F

(16) In a SLR, R^2 is equal to the square of the sample correlation between the observed Y and X values.

[✓ T] F

(17) The R^2 for a regression of Y onto X is the same as the R^2 for the regression of X onto Y.

[✓ T] F

(18) In SLR, $Corr(Y, \hat{Y})$ is always 1.

T [✓ F]

$corr(Y, \hat{Y}) = corr(Y, X) \cdot sign(b)$

(19) In SLR, $Corr(X, \hat{Y})$ is always 1.

T [✓ F]

slope can be negative

(20) In SLR, if the $R^2 > .8$, we are guaranteed to make an accurate, precise prediction.

T [✓ F]

11 Question

An investigator would like to survey a set of people in order to learn what fraction of them use illegal drugs. The survey involves the standard approach of randomly selecting a subset of individuals to survey from the complete list of population members.

The investigator is concerned that potential respondents will be reluctant to answer a question about drug usage truthfully.

The investigator will have each respondent flip a coin.
If it comes up tails, the respondent will answer 1 (yes) or 0 (no) to the question "Is the first digit of your social security number even".
If the coin comes up heads, the respondent will answer 1 (yes) or 0 (no) to the question "do you use illegal drugs".

Thus, each respondent will answer 1 or 0 (yes or no) but the investigator does not know which of the two questions the respondent is actually replying to. The hope is that since the investigator does not know which question was asked, the respondent will give the correct answer.

Let Q be the random variable which is 1 if the coin comes up heads (the drugs question is asked) and 0 if the coin comes up tails (the digit question is asked).

Let R be the random variable representing the answer (1 for yes, 0 for no).

Thus, $P(Q = 1) = P(Q = 0) = .5$.

The investigator believes that $P(R = 1 \mid Q = 0) = .5$.

The investigator would like to know $P(R = 1 \mid Q = 1)$.

Since we will use $P(R = 1 \mid Q = 1)$ a lot, to simplify notation let's also call it p_1:
$p_1 = P(R = 1 \mid Q = 1)$.

First, let's look at things from the point of the respondent.
He might wonder if the investigator can guess what question was asked. For example, if drug use is very low in the population, then a no answer might suggest it was the drug question that was asked.

To investigate this, a respondent supposes that prior to collecting the data, the investigator might believe $p_1 = .1$

11.1

Suppose $p_1 = .1$, what is $P(Q = 1, R = 1)$.

see sol.

11.2

Suppose $p_1 = .1$. what is $P(R = 1)$?

see sol.

11.3

Suppose $p_1 = .1$, what is $P(Q = 1 \mid R = 1)$?

see sol.

11.4

Suppose $p_1 = .1$, what is $P(Q = 1 \mid R = 0)$?

see sol.

11.5

How do the above probabilities make the respondent feel?
Can the investigator guess the question in a way that matters to the respondent?

see sol.

Now let's look at things from the point of view of the investigator.
He is trying to esitmate $p_1 = P(R = 1 \mid Q = 1)$.
The survey allows him to estimate $P(R = 1)$.
To simplify notation, let's call this p, $p = P(R = 1)$.

11.6

Note that given p_1, we can figure out p.

Write p as a linear function of p_1.

Note that you can check your function by plugging in $p_1 = .1$ and making sure you get the same answer as you got above!

see sol.

11.7

Now suppose the survey is done and 450 out of 1,000 respondents answer yes.
What is your estimate of p?

see sol.

11.8

Now suppose the survey is done and 450 out of 1,000 respondents answer yes.
What is your estimate of p_1?

see sol.

11.9

Is your estimator for p_1 unbiased?

see sol.

11.10

Now suppose the survey is done and 450 out of 1,000 respondents answer yes. Give a 95% confidence interval for p.

$$CI = \hat{p} \pm 2 \times SE$$

see sol. -> [.42, .48]

11.11

Now suppose the survey is done and 450 out of 1,000 respondents answer yes. Give a 95% confidence interval for p_1.

see solutions
plug CI from 11.10 into linear fct.

11.12

Now suppose the investigator cheated and observed the question.

It actually worked out the each question was asked 500 times and 250 of the digit questions were answered yes and 200 of the drug questions were answered yes.

Give this additional information, give a 95% confidence interval for p_1.

see sol.

11.13

Which p_1 confidence interval is smaller, the one where did not know about which question was asked or the one where you did?

Why does this make sense?

see sol.

Given $X_i = c$ means treat X_i as "$N(c, 0)$" for the usual formulas to apply.

12 Question

Suppose

$$Y = 20 + 5X_1 + 3X_2 + \epsilon, \;\; \epsilon \sim N(0, 25).$$

and $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$, X_1, X_2, ϵ all independent.

12.1

What is $E(Y)$?

$$20 + 5 \times 0 + 3 \times 0 + 0 = \underline{\underline{20}}$$

12.2

What is $Var(Y)$?

$$5^2 \times 1 + 3^2 \times 1 + 25 = \underline{\underline{59}}$$

12.3

Given $X_1 = 1$, what is $E(Y)$?

$$20 + 5 \times 1 + 3 \times 0 + 0 = \underline{\underline{25}}$$

12.4

Given $X_1 = 1$, what is $Var(Y)$?

$$3^2 \times 1 + 25 = \underline{\underline{34}}$$

12.5

Given $X_1 = 1$, $X_2 = -1$, what is $E(Y)$?

$$20 + 5 \times 1 + 3 \times (-1) + 0 = \underline{\underline{22}}$$

12.6

Given $X_1 = 1$, $X_2 = -1$, what is $Var(Y)$?

$$\underline{\underline{25}}$$